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 g with the property that $g(t) = tg(1/t)$ for all $t > 0$ and $g(0) = 0$; examples include $g(t) = \sqrt{t}$ and $g(t) = \min\{t, 1\}$ 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

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

Let $T_{p,X\to Y} = \infty$ on the rest of M. Thus, by our choice of g, 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, $\mathrm{supp}(p) = \{X \mid p(X) > 0\}$, 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 C.1 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.1. 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.8 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.9): after a certain length, longer and longer sequences are sufficiently less likely under p. This assumption holds for some models and not others (Section B). In Propositions A.10 and A.11 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.12. We provide examples of kernels that satisfy all our required assumptions in Section C.2. 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.9 and k is a deltable vector field kernel obeying Assumption A.12 A or a deltable kernel on S obeying Assumption A.12 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 $\text{KSD-B}_{p,k}(q_n) \to 0$.

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 $A(X) = \sum_{YMX} T_{p,Y\to X}\sqrt{k((X,Y),(X,Y))}$. If $\sum_X |p(X) - q_n(X)|A(X) \to 0$ then* $\text{KSD-B}_{p,k}(q_n) \to 0$.

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.

Stewart N Ethier and Thomas G Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, September 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.

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 for properties of KSDs

In this appendix we prove our assertions made in the main text. In section A.1 we lay out our notation.

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 $\mathrm{supp}(p) = \{X \mid p(X) > 0\}$ and $M_{pp} = \{(X, Y) \in M \mid X, Y \in \mathrm{supp}(p)\}$.

We define the set of bounded functions on S $C_b(S)$ 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 real sequences $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$, 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 $a_n \leq C b_n$. We write $a_n \sim b_n$ when $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. 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 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 $T : H \to \mathbb{R}$. The function $X, Y \mapsto T(X)k(X, Y)T(Y)$ is a kernel on H we will denote with a superscript as such: k^T.

A.2 Scalar and vector field KSDs and their computability

First we prove that scalar field KSDs can be understood as an instance of our vector field KSDs.

Proposition A.1. *For a symmetric non-negative definite 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 $f \in \mathcal{H}_{k^\nabla}$,

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

Finally, One can check similar to the proof of proposition A.2 that

$$\begin{aligned}
&\sup_{\|f\|_k \leq 1} E_q \mathcal{T}_p \nabla f \\
&\quad = E_{X,X' \sim q} \sum_{YMX, Y'MX'} T_{p,X \to Y} T_{p,X' \to Y'} \left(k(Y, Y') - k(X, Y') - k(Y, X') + k(X, X')\right) \\
&\quad = E_{X,X' \sim q} \sum_{YMX, Y'MX'} T_{p,X \to Y} T_{p,X' \to Y'} k^\nabla((X, Y), (X', Y')) \\
&\quad = \sup_{\|f\|_{k^\nabla} \leq 1} E_q \mathcal{T}_p f.
\end{aligned}$$

□

We now prove proposition 2.1 along with another form of the KSD-B that will be useful later.

Proposition A.2. *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

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

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, KSD-B$_{p,k}(q)^2 = \|\phi_k\|_k^2$.

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

Note that since all quantities in the expectation and sum are positive, equation 4 shows the absolute integrability of the expectation and sum. Thus we can rearrange terms to get

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

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 selecting the kernel: note that $q(X)T_{p,Y\to X} \geq 1$, 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.3 Stochastic processes on sequences

Before proceeding it is important to note the connection between the Stein operator and a stochastic process on S. We will make some assertions that we will use to prove later results.

Let p be a distribution on S. Then $\mathcal{L}_p = \mathcal{T}_p \nabla$ is an operator on the space of functions on S. It can be represented as a matrix with entry X, Y equal to $(\mathcal{L}_p \delta_Y)(X)$. This is $T_{p,X\to Y} > 0$ if $X \neq Y$ and is $-T_p(X) = -\sum_{YMX} T_{p,X\to Y}$ if $X = Y$. In particular, this matrix is a so-called Q-matrix and thus defines a stochastic process on S [Liggett, 2010]. We assume this process is non-explosive. Call this process $(X_t)_t$. Define the distribution of X_t conditional on $X_0 = X$ as $P_t(X)$. Now define the semi-group operators $P_t f(X) = E[f(X_t)|X_0 = X]$. Thus, if $f \in C_b(S)$, $\|P_t f\|_\infty \leq \|f\|_\infty$. As well, $P_t f(X)$ is continuously differentiable in t and $\frac{d}{dt} P_t f = \mathcal{L} P_t f(X) = P_t \mathcal{L} f(X)$. Finally, we have that if $E_q P_t f = E_q f$ for all $f \in C(S)$, then $q = p$.

Proposition A.3. *Say $\phi : S \to [0,\infty)$ such that $\mathcal{L}\phi(X) \leq 0$ for all but finitely many $X \in S$. Then $P_t\phi(X) < \infty$ for all $X \in S$ and $\frac{d}{dt} P_t \phi(X) = \mathcal{L} P_t \phi(X)$.*

Proof. Call, for any positive function ϕ, $G(\phi, 0, X, t) = 0$ and

$$G(\phi, n, X, t) = e^{-tT_p(X)}\phi(X) + \int_0^t ds e^{-T_p(X)(t-s)} \sum_{YMX} T_{p,X\to Y} G(\phi, n-1, X, s). \tag{5}$$

By induction, it is easy to see that $G(\phi, n, X, t)$ is increasing in n and, since the right hand side is positive, by absolute integrability

$$G(\phi, n, X, t) = \sum_Y G(\delta_Y, n, X, t)\phi(Y).$$

Note by theorem 2.25 of Liggett [2010], $G(\delta_Y, n, X, t)$ increases to $P_t(X)(Y)$ as $n \to \infty$, so by monotone convergence,

$$G(\phi, n, X, t) \to P_t\phi(X).$$

If this quantity is finite for all X, $\frac{d}{dt}P_t\phi(X) = \mathcal{L}P_t\phi(X)$ then follows by taking limits of equation 5, noting that this implies that $P_t\phi$ is continuous in t, and then differentiating with respect to t. Thus the theorem is proven if we show that $\sup_n G(\phi, n, X, t) < \infty$.

Say $G(\phi, n, X, t) \leq \phi(X) + C_t$ for some large C_t. Then

$$\begin{aligned}
G(\phi, n+1, X, t) \leq & e^{-tT_p(X)}\phi(X) + \int_0^t ds e^{-T_p(X)(t-s)} \sum_{YMX} T_{p,X\to Y}\left(\phi(X) + C_t\right) \\
= & e^{-tT_p(X)}\phi(X) + \int_0^t ds e^{-T_p(X)(t-s)} \left(T_p(X)\left(\phi(X) + C_t\right) + \mathcal{L}\phi(X)\right) \\
= & e^{-tT_p(X)}\phi(X) + \left(1 - e^{-tT_p(X)}\right)\left(\phi(X) + C_t\right) + \frac{1 - e^{-tT_p(X)}}{T_p(X)}\mathcal{L}\phi(X) \\
= & \phi(X) + C_t + \left(\frac{1 - e^{-tT_p(X)}}{T_p(X)}\mathcal{L}\phi(X) - e^{-tT_p(X)}C_t\right).
\end{aligned}$$

Since $\mathcal{L}\phi$ is positive on only finitely many X, C_t can be chosen to make the quantity in brackets negative for all X. Thus by induction, $\sup_n G(\phi, n, X, t) < \phi(X) + C_t$ for all X, t. □

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. Say $V : S \to [1, \infty)$ is a function such that $V(X) \to \infty$ as $|X| \to \infty$. $\mathcal{L}_p V \leq K - \varphi \circ V$ 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*

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

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 super-martingale. To see this, first note $\varphi \circ V \lesssim V$, so $P_t\varphi \circ V < \infty$. Now if $t, s \geq 0$,

$$\begin{aligned}
E\left[V(X(t+s)) - V(X(s))|X(s)\right] = & P_t V(X(s)) - V(X(s)) \\
= & \int_0^t dr \mathcal{L}P_r V(X(s)) \\
= & \int_0^t dr P_r \mathcal{L}V(X(s)) \\
\leq & \int_0^t dr \left(K - P_r\varphi \circ V(X(s))\right) \\
= & E\left[\int_s^{t+s} dr \left(K - \varphi \circ V(X(r))\right)|X(s)\right].
\end{aligned}$$

□

A.4 Faithfulness and deltability

We now prove the faithfulness of KSDs. To do so, we will need to make an assumption that guarantees that $\mathcal{H}_k$ is large enough.

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.*

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 $C_C(S) \subset \mathcal{H}_k$. Thus 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].

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 thus asks whether this implies $q = p$. This is obviously true for $f \in \mathcal{T}_p(C_{C,vf}(M))$ and the proof will simply follow from 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 for a scalar field kernel k. Instead we will notice that $\mathcal{T}_p\nabla$ is the infinitesimal generator for a stochastic process on S which has a unique stationary distribution p. Unfortunately whether $E_q\mathcal{T}_p\nabla f = 0$ for all $f \in C_C(S)$ implies $q = p$ is a delicate question. The veracity of the implication can depend on connectivity the sample space S and whether we allow q to be signed or non-finite. One sufficient condition for this implication to hold is that $C_C(S)$ is a core for the unbounded operator $\mathcal{T}_p\nabla$ in $L^1(q)$ (proposition 9.2 of Ethier and Kurtz [2009]). Define $\text{flux}_p(X) = \sum_{YMX} T_{p,X\to Y}$. The integrability condition $E_q\text{flux}_p < \infty$ is sufficient for this as we show below.

To prove faithfulness, we will look at this next lemma first.

Lemma A.6. *Say p has connected support and q is a distribution on S. If $E_q\mathcal{T}_p f = 0$ for all $f \in C_{C,vf}(M)$ or $E_q\text{flux}_p < \infty$ and $E_q\mathcal{T}_p\nabla f = 0$ for all $f \in C_C(S)$ then $q = p$.*

Proof. For the first claim, first note that 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 this case, $E_q\mathcal{T}_p\delta_{(X,Y)} = \infty$, a contradiction. Thus $\text{supp}(q) \subseteq \text{supp}(p)$. Now by equation 2[1] we have $q(X)/p(X) = q(Y)/p(Y)$ for all $X, Y \in \text{supp}(p)$; since the support of p in connected, this implies that $q = p$.

For the second claim, note $E_q\text{flux}_p < \infty$ implies $\text{supp}(q) \subseteq \text{supp}(p)$. Now recall, as described above, $\mathcal{L} = \mathcal{T}_p\nabla$ is an infinitesimal generator for a semi-group $(P_t)_{t\geq 0}$. We also have, for any $f \in C_C(S)$, $\mathcal{L}f \in C_C(S)$, $\|P_t f\|_\infty \leq \|f\|_\infty$, and for all $X \in S$, $\frac{d}{dt}P_t f(X) = \mathcal{L}P_t f(X) = P_t\mathcal{L}f(X)$. Define the measure, for $A \subset S$, $q_t(A) = E_{X\sim q}P_t(X, A) = E_{X\sim q}P_t\mathbb{1}_A(X)$. If $q_t = q$ for some $t > 0$ then q is an invariant measure for P_t, so, $q = p$. Let $Z \in S$. As $P_r\mathcal{L}\delta_Z$ is uniformly bounded for all r, we can swap integration and differentiation to arrive at $\frac{d}{dt}q_t(Z)|_{t=s} = E_q\mathcal{L}P_t\delta_Z$. If there is a sequence of $g_n \in C_C(S)$ such that $E_q|\mathcal{L}g_n - \mathcal{L}P_t\delta_Z| \to 0$, then $E_q\mathcal{L}P_t\delta_Z = \lim_n E_q\mathcal{L}g_n = 0$, so that $q_t(Z) = q_0(Z) = q(Z)$ for all t and the claim follows.

We will now construct this sequence. Pick a sequence of subsets of S, $\Lambda_1, \Lambda_2, \ldots$ with $\cup_{n=1}^\infty \Lambda_n = S$ and define $g_n = \mathbb{1}_{\Lambda_n}P_t\delta_Z$. Define $\epsilon_n = |\mathcal{L}g_n - \mathcal{L}P_t\delta_Z|$. Since $\|P_t\delta_Z\|_\infty \leq 1$, $\epsilon_n(X) \leq \sum_{YMX} T_{p,X\to Y}$ for all $X \in S$. As well, $\epsilon_n(X) = 0$ for any $X \in \Lambda_n$ such that $Y \in \Lambda_n$ for all YMX. Thus, by dominated convergence, since $\cup_{n=1}^\infty \Lambda_n = S$, $E_q\epsilon_n \to 0$. □

Now we use this lemma to show detection of tight non-convergence of the KSD. This will imply proposition 3.1.

Proposition A.7. *Say supp(p) is connected and $(q_n)_n$ is a tight sequence of distributions on S satisfying $\text{KSD-B}_{p,k}(q) \to 0$. If k is a deltable vector field kernel or k is a deltable kernel on S 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 (B) and $E_q\text{flux}_p < \infty$ $\text{KSD-B}_{p,k}(q) = 0$ only if $p = q$.*

Proof. Assume k is a deltable vector field kernel. Say $\text{KSD-B}_{p,k}(q) \to 0$ 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

[1] We do not need p, k-integrability as $\mathcal{T}_p f$ is nonzero on only finitely many point for all $f \in C_{C,vf}(M)$.

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.6, $q = p$, a contradiction.

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

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

□

A.5 Detection of non-convergence

We will now prove results that describe conditions under which the KSD-B detect convergence and non-convergence.

In the next proposition we will show an example of a distribution p for which the KSD-B does not detect non-convergence.

Proposition A.8. *Let $p(X) \propto |\mathcal{B}|^{-L} e^{-\mu L}$ 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 $KSD_{k,p}(q_n) \to 0$ 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 2 are non-zero only for $(X,Y) \in N_L$. Thus,

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

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

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

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

The additional assumptions we make on p will essentially ask that it has uniformly decreasing tails. First define the quantities

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

Define all quantities to be ∞ if no $X \in \text{supp}(p)$ has $|X| = L$. $\text{del}_p^-(L)$ describes a minimum propensity to gain a deletion, $\text{ins}_p^+(L)$ a maximum propensity to gain an insertion. We phrase our assumption of uniformly decreasing tails as

Assumption A.9. *We assume 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 $\epsilon_V > 0$.

V_p is our Foster-Lyapunov function that will allow us to determine the convergence of P_t to p [Hairer, 2021]. This property only asks that $\mathrm{gap}_p(L)$ increases quickly enough. To put this into a more interpretable form, note that the since V_p is convex and goes to ∞, the right hand side is eventually less than $(\log V_p)'(L)^{-1} V_p^{-\frac{1}{2+\epsilon_V}}(L)$. For $V_p = L^\alpha$ for some $0 < \alpha < 1$ this quantity is $L^{1-\alpha\frac{1}{2+\epsilon_V}}$. Another option is $V_p(L) = (\log(L))^\beta$ for some $\beta > 0$, in which case the above quantity is $L\log(L)^{1-\frac{\beta}{2+\epsilon_V}}$. Thus this assumption is implied by the assumptions that $\mathrm{supp}(p)$ be finite or $\mathrm{gap}_p(L) \gtrsim L^\alpha$ for some $\alpha > 1/2$. In general, the faster gap_p increases, the slower we can make V_p increase. Note that this insertion necessitates that if $X \notin \mathrm{supp}(p)$ and Y can be reached from X via insertions then $Y \notin \mathrm{supp}(p)$. In particular, it implies that $\mathrm{supp}(p)$ is connected.

In the next propositions we show examples where p has uniformly decreasing tails but the KSD-B cannot detect non-convergence if k is a bounded scalar field kernel or k is a vector field kernel with thin tails.

Proposition A.10. *There is a distribution p on S such that $\mathrm{gap}_p(L) \sim L$ 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 some distribution supported on $\{\emptyset, A, AA, AAA, \dots\}$ for $A \in \mathcal{B}$, calling $p(L) = p(L \times A)$ for any number L. Assume for now that p is decreasing in L and define $r_L = g\left(\frac{p(L)}{p(L-1)}\right) < 1$ and $\tilde{r}_L = g\left(\frac{p(L-1)}{p(L)}\right) < 1$ (with $r_0 = \infty$). Thus, $r_L \leq 1 \leq \tilde{r}_L$ 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 $\|f\|_\infty \leq 1$,

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

Let $\tilde{q}_{m,n}(L) = \frac{m\tilde{r}_m}{L\tilde{r}_L}$ for $m \leq L \leq n$ and $\tilde{q}_{m,n}(L) = 0$ for $L > n$ and $L < m$. Note that we then have that $\frac{\tilde{q}_{m,n}(L+1)}{\tilde{q}_{m,n}(L)} = \frac{L\tilde{r}_L}{(L+1)\tilde{r}_{L+1}}$. Now let $q_{m,n} = \tilde{q}_{m,n}/Z_{m,n}$ where $Z_{m,n} = \sum_{l=m}^{n} \tilde{q}_{m,n}$. We

have that $q(L)L\tilde{r}_L = \frac{m\tilde{r}_m}{Z_{m,n}}$ for all $m \leq L \leq n$. Thus,

$$
\begin{aligned}
E_{q_{m,n}}\mathcal{T}_p\nabla f =& f(m-1)\frac{m\tilde{r}_m}{Z_{m,n}} - f(n)\frac{m\tilde{r}_m}{Z_{m,n}} \\
&+ \sum_{L=m+1}^{n} f(L)\left(q_{m,n}(L-1)Lr_L - q_{m,n}(L)(L+1)r_{L+1}\right) \\
&- f(m)q_{m,n}(m)(m+1)r_{m+1} + f(n+1)q_{m,n}(n)(n+1)r_{n+1} \\
=& \frac{m\tilde{r}_m}{Z_{m,n}}\Bigg(f(m-1) - f(n) \\
&- f(m)\frac{m+1}{m}\tilde{r}_m^{-1}r_{m+1} + f(n+1)\frac{n+1}{n}\tilde{r}_n^{-1}r_{n+1}\Bigg) \\
&+ \sum_{L=m+1}^{n} q_{m,n}(L-1)f(L)Lr_L\left(1 - \frac{(L+1)(L-1)}{L^2}\frac{r_{L+1}\tilde{r}_{L-1}}{\tilde{r}_L r_L}\right) \\
\leq& \frac{6m\tilde{r}_m}{Z_{m,n}} + \sup_{L>m} Lr_L\left|1 - \left(1 - \frac{1}{L^2}\right)\frac{r_{L+1}\tilde{r}_{L-1}}{\tilde{r}_L r_L}\right|.
\end{aligned}
\tag{6}
$$

Now pick p to be the distribution with $p(L) \propto= 2^{-L}$, so, $\tilde{r}_L, r_L$ are fixed constants $\tilde{r} > r$. Note that for any fixed m,

$$\frac{1}{m\tilde{r}_m}Z_{m,n} = \sum_{l=m}^{n}\frac{1}{L\tilde{r}} \to \infty \text{ as } n \to \infty.$$

Thus the first term of equation 6 goes to 0. We also have that

$$\frac{r_{L+1}\tilde{r}_{L-1}}{\tilde{r}_L r_L} = \frac{r\tilde{r}}{\tilde{r}r} = 1.$$

Thus,

$$Lr_L\left|1 - \left(1 - \frac{1}{L^2}\right)\frac{r_{L+1}\tilde{r}_{L-1}}{\tilde{r}_L r_L}\right| = Lr\frac{1}{L} \to \quad 0 \text{ as } L \to \infty.$$

Thus, by sending $m, n \to \infty$, $\sup_{\|f\|_\infty \leq 1} E_{q_{m,n}}\mathcal{T}_p\nabla f \to 0$ while $q_{m,n} \not\to p$.

On the other hand,

$$\text{gap}_p(L) = L\tilde{r} - (L+1)r \sim L.$$

□

This next result is similar in idea to the example of theorem 6 of Gorham and Mackey [2017].

Proposition A.11. *Let $p(X) \propto e^{-\lambda|X|}|\mathcal{B}|^{-|X|}$ and k be a 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-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 $g\left(e^{\mu}|\mathcal{B}|\right) = c$, $T_{p,X\to Y} \leq c(|X|+1)$. For distinct points $X_1, \ldots, X_N \in \mathcal{B}^L$ 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,

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

We will now pick, for each L, a set of sequences $X_1, \ldots, X_{N_L} \in \mathcal{B}^L$ such that $R_L = \min_{n \neq m} d_H(X_n, X_m)$ has $L^4 \left(N^{-1} + R^{-(4+\epsilon)} \right) \to 0$ and the result follows. Say $X_1, \ldots, X_{N_L} \in \mathcal{B}^L$ is a maximal set of sequences with $d_H(X_n, X_m) > R_L = L|\mathcal{B}|/20$ for all $n \neq m$. 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 $(X_n)_n$ would not be maximal. 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 $|B(X_1, R_L)|/|\mathcal{B}|^L = P(Z \geq R_L)$. On the other hand, calling $t = \log\left(\frac{R_L}{L|\mathcal{B}|}\right) = -\log 20$,

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

Thus, $N_L \geq e^{\frac{1}{2}L|\mathcal{B}|}$. so that $L^4 \left(N_L^{-1} + R_L^{-(4+\epsilon)} \right) \to 0$ as $L \to \infty$. □

We thus also assume that there is a function in the RKHS of the kernel that has thick enough tails, which of course necessitates that k itself has thick tails.

Assumption A.12. *Say p is a distribution on S that satisfies assumption A.9 with V_p. (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*

$$\sum_L \frac{\inf_{|X|=L} \mathcal{T}_p \tilde{f}(X)}{\text{ins}_p^+(L) V_p(L+1)} = \infty. \tag{7}$$

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

$$\sum_L C_L \wedge C_{L+1} = \infty \text{ where } C_L = \frac{\inf_{|X|=L} \mathcal{T}_p \nabla \tilde{f}(X)}{\left(\sup_{|X|=L} \text{flux}_p(X) \right) V_p(L+1)}. \tag{8}$$

The later assumption is very similar to the first with the exceptions of 1) the operator $\mathcal{T}_p \nabla$ instead of $\mathcal{T}_p$, 2) replacing the ins_p^+ terms with a possibly much larger flux_p term, and 3) the sum is over minima of sequential terms. The last condition 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.13. *Say p is a distribution on S obeying assumption A.9. If $g \in C_b(S)$ there is a $f_g : S \to \mathbb{R}$ such that $\mathcal{T}_p \nabla f_g = g$ and $f_g(X) \leq CV_p(X)\|g\|_\infty$ for a universal constant C.*

Proof. $\mathcal{L} = \mathcal{T}_p \nabla$ 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 $|X| = L$,

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

By our assumptions on V_p, the first term is negative and $\mathrm{gap}_p(L)\Delta V_{p,L} \gtrsim \varphi(V_p(L-1))$ where $\varphi(x) = x^{(1+\epsilon)/(2+\epsilon)}$.. Thus there are constants C_1, C_2 such that for all $X \in S$

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

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

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

Thus the first term of equation A.5 is eventually negative.

Now assume $g \in C(S)$. 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

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

with $|f_g|(X) \leq C'\|g\|_\infty V_p(X)$. By absolute integrability, we can also write

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

□

Theorem A.14. *Say p is a distribution on S obeying assumption A.9 and k is a deltable vector field kernel obeying assumption A.12 A or a deltable kernel on S obeying assumption A.12 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. Let $g \in C_b(S)$ with $\|g\|_\infty \leq 1$, so by lemma A.13, 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.12 A. Let $\tilde{f} \in \mathcal{H}_k$ satisfy equation 7 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

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

Note

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

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

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

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 $\sup_L \Delta v_L^{\epsilon,L'} R_L = \epsilon$ 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.12 B. The proof is very similar. Let $\tilde{f}\in\mathcal{H}_k$ satisfy equation 8 and have $\mathcal{T}_p\nabla\tilde{f}(X)\to\infty$ as $|X|\to\infty$. There is thus a $\zeta\in\mathbb{R}$ such that $\mathcal{T}_p\nabla\tilde{f}(X)+\zeta > 0$ for all $X\in S$. 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

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

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

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

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 $v_L = v_{L-1} - \epsilon R_{L-1}^{-1} \wedge R_L^{-1} \wedge (v_{L-1})$ for $l \geq L$. Thus $\tilde{\Delta} v_L \leq \epsilon R_L^{-1}$. By assumption $\sum_L R_L^{-1} \wedge R_{L+1}^{-1} = \infty$ so $v^{\epsilon,L'}$ is eventually 0. We thus have $\sup_L \tilde{\Delta} v_L^{\epsilon,L'} R_L = \epsilon$ 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 $\epsilon \to 0$. □

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

Proposition A.15. *Say k is a vector field kernel and $p, q_1, q_2, \ldots$ are p, k-integrable distributions on S. Call $A(X) = \sum_{YMX} T_{p,Y\to X}\sqrt{k((X,Y),(X,Y))}$.*

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

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

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

Which proves the result. □

A.6 Efficient approximate kernelized Stein discrepancies

In this section we will prove proposition **??**.

Proposition A.16. *Let p be a distribution on S, $(q_n)_n$ a sequence of distributions on S with $\sup_n E_{q_n}\text{flux}_p < \infty$ and k is a bounded vector field kernel. For each n, let $(Y_{X,n,m})_{m=1}^{M_{X,n}}$ be iid samples from $N(X)$ sampled with probabilities $(T_{p,X\to Y}/T_p(X))_{Y\in N(X)}$. Define the approximate KSD*

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

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

Proof. Let $M_{X,n}$ be a family of number such that $M_{X,n} \geq C|X|n(\log n)^2$ for some $C > 0$. Call $p(Y|X) = T_{p,X\to Y}/T_p(X)$. Sample $(Y_{X,n,m})_{m=1}^{M_{X,n}}$ iid from $p(Y|X)$ for all $X \in S$. Call $\hat{p}_n(Y|X) = \frac{1}{M_{X,n}}\sum_{m=1}^{M_{X,n}} \delta_{Y_{X,n,m}}$. Note that if the later quanitty is finite, by the same proof of proposition 2.1,

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

Note also that since k is bounded, $(q_n)_n$ are p, k-integrable. Thus, since $\|f\|_k \leq 1$ implies that $\|f\|_\infty \leq 1$,

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

where $\|\cdot\|_{\text{TV}}$ is the total variation metric. We will show that

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

as $n \to \infty$ almost surely, which will finish the proof.

Pick a sequence of positive numbers $\epsilon_1, \epsilon_2, \ldots$. Call $\mathcal{F}_X = \{f : N(X) \to \{-1, 1\}\}$ so

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

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,

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

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 can actually set the number of samples to be much smaller if we assume q is finite.

Proposition A.17. *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 n, let $(Y_{i,n,m})_{m=1}^{M_{i,n}}$ be iid samples from $N(X)$ sampled with probabilities $(T_{p,X\to Y}/T_p(X))_{Y\in N(X)}$. Define the approximate KSD*

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

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

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

Proof. Let $M_{i,n}$ be a family of number such that $M_{i,n} \geq C\log(L_n)(\log n)$. For each n, let $(Y_{i,n,m})_{m=1}^{M_{i,n}}$ be iid samples from $N(X)$ sampled with probabilities $(T_{p,X\to Y}/T_p(X))_{Y\in N(X)}$. 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.16,

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

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation metric. 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.

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 $n \to \infty$,

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

□

B Distributions and their uniform tails

In this section we will consider some examples of distributions that do and do not satisfy assumption A.9.

First let's look at an auto-regressive model that does not satisfy assumption A.9.

Proposition B.1. *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 $p(A|AA) = 0.1$, $p(B|AA) = 0.8$, $p(\$|AA) = 0.1$, and $p(A|AB) = 0.8$, $p(B|AB) = 0.1$, $p(\$|AB) = 0.1$, $p(A|BA) = 0.8$, $p(B|BA) = 0.1$, $p(\$|BA) = 0.1$ where \$ represent the end of the sequence. Then p does not satisfy assumption A.9.*

Proof. $p(L \times A) = 0.1^{L-1}$, so $\mathrm{del}^-(L) \leq Lg(0.1)$. However, $p(L_1 \times A + B + L_2 \times A) = 0.1^{L_1+L_2-2-1}0.8^3$, so, if $L_1 + L_2 = L$, $p(L_1 \times A + B + L_2 \times A)/p(L \times A) = 0.8^3 0.1^{-2}$, for $\mathrm{ins}^+(L) \geq (L+1)g(0.8^3 0.1^{-2}) \geq \mathrm{del}^-(L)$. □

On the other hand there are obvious examples of distributions that do satisfy assumption A.9, 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 $\mathrm{gap}_p(L) \sim L^2$.

Proposition B.2.

Proof. Note for every sequence X and sequence of hidden states Z,

$$\sum_{|Y|=L+1, XMY} T_{p,X\to Y} = \frac{1}{p(X)} \sum_{l=0}^{L} \sum_{Y \mathrm{ins}_l X} p(Y)$$

where the sum $\sum_{Y \mathrm{ins}_l X}$ is over all $|\mathcal{B}|$ insertions of X at position l.

$$\sum_{Y \mathrm{ins}_l X} p(Y) = \sum_Z p(Z) \prod_{l'=0}^{l-1} p(X_{l'}|Z_{l'}) \prod_{l'=l+1}^{L+1} p(X_{l'}|Z_{l'}).$$

If position l is an insertion in Z, call $\tilde{Z}$ the same hidden states without l. In this case $p(Z)/p(\tilde{Z}) \leq e^{-\mu^-}$ and $Z \mapsto \tilde{Z}$ is an injection (in fact, a bijection). If position l is an aligned letter in Z, call $\hat{Z}$ the same hidden states . In this case $p(Z)/p(\tilde{Z}) \leq$ and $Z \mapsto \hat{Z}$ is an injection.

Then, calling $Q_{L,l} = p(Z_l$ not an insertion), we have

$$\sum_{Y \mathrm{ins}_l X} p(Y) \leq Q_{L,l} \sum_{\hat{Z}} p(\hat{Z})p(X|\hat{Z}) + e^{-\mu^-} \sum_{\tilde{Z}} p(\tilde{Z})p(X|\tilde{Z}) = Q_{L,l} + e^{-\mu^-}p(X).$$

□

C Kernels

We've described above requirements on a kernel for good behaviour from a KSD. Now we describe scalar and how to build kernels that satisfy these requirements.

In the next sections we will describe 1) how to build vector field kernels, and 2) simple example kernels that satisfies the above requirements.

C.1 Vector field kernels for KSDs

We now describe how to build vector field kernels. We start by noting a correspondence between kernels without algebraic restrictions on some space and vector field kernels. To state this correspondence, we will need the following definition.

Definition C.1. *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 $M^\sigma = \sigma^{-1}(1)$. For a $(X, Y) \in M$, define $(X, Y)^\sigma = (X, Y)$ if $\sigma(X, Y) = 1$ and (Y, X) otherwise.*

Proposition C.2. *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 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. A kernel k on M^σ such that $\delta_{(X,Y)} \in \mathcal{H}_k$ for all $(X, Y) \in M^\sigma$ corresponds to a deltable vector field kernel.

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

Proposition C.3. *Let k, k' be kernels on S. The following are kernels on M^σ.*

$$((X, Y), (X', Y')) \mapsto k(X, X')k'(Y, Y')$$
$$((X, Y), (X', Y')) \mapsto (k(X, X') + k'(Y, Y'))^2$$
$$((X, Y), (X', Y')) \mapsto k(X + Y, X' + Y')$$
$$((X, Y), (X', Y')) \mapsto k(X, X').$$

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

C.2 Coercive vector field kernels with delta functions

We will in this section give an example of a deltable vector field kernel obeying assumption A.12 A and an example of a deltable kernel on S obeying assumption A.12 B.

Define the inverse multiquadratic Hamming kernel 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 \$.

To define our scalar kernel, we define

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

for $A(X) = (|X| + 1)^{3/2}$. By an unpublished result, this is a well defined deltable kernel. If p is a pHMM, p is p, k-integrable. As well, $k_\emptyset(X) = |X| + 1$, so since $\mathrm{ins}_p^+(L) \lesssim Le^{-\mu^-}$ and $\mathrm{del}_p^-(L) \gtrsim Le^{\mu^-}$, we have $V_p(L) \gtrsim \log(L)^2$ and $\mathcal{T}_p k_\emptyset(X) \gtrsim \mathrm{gap}_p(|X|) \sim |X|$. Thus, k satisfies assumption A.12 B.

Finally, define the vector field kernel

$$k((X, Y), (X', Y')) = \delta_X(X') + k_H(X, X')$$

for $(X, Y), (X', Y') \in M^\sigma$ for a proper sign σ. k is bounded and deltable. This time define $\tilde{f} = \sum_{|Y|=1} k_{(\emptyset,Y)}$. In this case, $\mathcal{T}_p \tilde{f}(X) \sim \mathrm{gap}_p(|X|)/\sqrt{|X|} \sim \sqrt{|X|}$. Thus, k satisfies assumption A.12 A.

C.3 Kernel proofs

Proposition C.4. *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 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. A kernel k on M^σ such that $\delta_{(X,Y)} \in \mathcal{H}_k$ for all $(X,Y) \in M^\sigma$ corresponds to a deltable vector field kernel.

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, distinct $(Z_n)_{n=1}^N \subset M$, 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.

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

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 also takes f to the above defined $\tilde{f}$ and can be extended to all of $\mathcal{H}_k$ to obey the same property. □

Proposition C.5. *Let k, k' be kernels on S. The following are kernels on M^σ.*

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

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

Proof. That these all define kernels is clear with some application of the Schur product theorem. The second statement relies on unpublished results. □