Research statement

Vu Dinh, Fred Hutchinson Cancer Research Center

My research focuses on phylogenetics and applied probability/statistics, with an emphasis on the development of next-generation Markov chain Monte Carlo (MCMC) methods for phylogenetic inference. I am also interested in computational methods for experimental design and control of biological systems, as well as machine learning algorithms and their applications in applied sciences.

Modern phylogenetic inference methods for modern data sets

Phylogenetics, the inference of evolutionary trees from molecular sequence data such as DNA, is an important enterprise enabling an evolutionary understanding of biological systems. Modern data sets in phylogenetics are typically large, heterogeneous and increasingly dynamic; however, current computational methods have not been able to handle such increasing complexities. My research aim is to extend our understanding about the essential objects involved in phylogenetics, to establish theoretical foundations for statistical analyses on tree spaces, and to use such knowledge to design fundamentally new inference methodologies.

Theoretical foundations for statistics on tree spaces: geometry of tree spaces and the phylogenetic likelihood surfaces. The set of phylogenetic trees forms a space with discrete (graph structure) and continuous (branch lengths) components. Most statistical methods are not developed with such spaces in mind. A central theme of my research is to extend statistical methods to more complex spaces, here the cubical complex model of tree space (Figure 1).

Geometric properties of phylogenetic likelihood surfaces play an essential role in analyses and designs of phylogenetic algorithms. Theoretical and simulation results indicate that the phylogenetic likelihood surface might be quite complex. This is further supported by one of our recent work [1], in which *we prove that one-dimensional phylogenetic likelihood functions may take the shape of any given arbitrary continuous function*. This analysis also helps develop specialized surrogate functions for branch length inference, which has been partially implemented in our open-source library for phylogenetic curve-fitting[1].

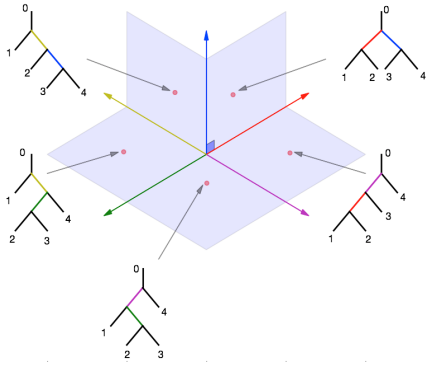

Figure 1: Tree space modeled as a (non-differentiable) cubical complex with convex geodesics.

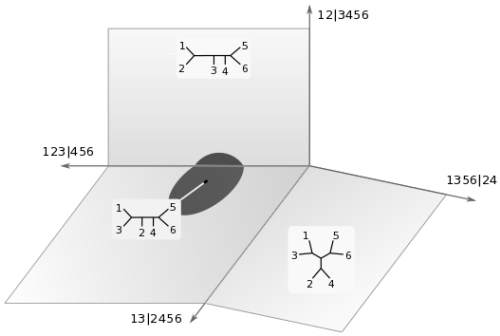

Figure 2: The ℓ_2-regularized estimator allows for using confidence regions on tree spaces to assess the support of topological properties even when data is not well-informed.

Phylogenetic regularization. Regularization has not yet had the impact on phylogenetics that it has in the rest of statistics. *We develop the first regularized estimator for tree reconstruction, which uses the squared geodesic distance on tree space as the penalty to derive an ℓ_2-type penalty.* This estimator incorporates information about the species tree to enhance the accuracy and stability of individual gene trees estimation [2]. We prove that this method is consistent, and derived its global convergence rate for estimating the discrete gene tree structure and continuous edge lengths simultaneously. Through analyses of the

[1] https://github.com/matsengrp/lcfit

phylogenetic likelihood surfaces and by the convexity of the geodesic distance, we prove that the estimator is adaptively fast convergent[2].

Exploring the cubical complex of phylogenetic trees with Hamiltonian dynamics. Hamiltonian Monte Carlo (HMC) is a powerful sampling algorithm which has been shown to outperform many existing MCMC algorithms in various contexts. However, the construction of an HMC sampling method for phylogenetic inference has been hindered by the discrete nature of the inference problem. To resolve this issue, *we develop* phyloHMC, *a probabilistic version of HMC on the cubical complex of phylogenetic trees*, and establish that the new integrator retains the good theoretical properties of Hamiltonian dynamics in classical settings[3]. We then prove that the resulting Markov chain is ergodic, and that the algorithm is capable of exploring the tree spaces more efficiently than traditional MCMC methods [3].

Online phylogenetic inference with Sequential Monte Carlo. Phylogenetics is being used in new dynamic ways, with sequence data continually being generated. This information needs to be be quickly analyzed by automated algorithms and presented for analysis. While appropriate computational infrastructure now exists, there are no phylogenetic algorithms for such a stream of data.

In a recent work, we develop the first online algorithm for phylogenetics: an online sequential Monte Carlo (OPSMC) method that continually updates phylogenetic posteriors given additional data [4]. *We derive the first set of bounds describing how phylogenetic likelihood surfaces change when new sequences are added*. These bounds enable us to characterize the theoretical performance of our sampler by bounding the effective sample size with a given number of particles and prove that for well-designed phylogenetic proposals, the diversity of OPSMC does not degenerate even as the problem dimension increases.

Future directions:

1. The relation between the geometry of the likelihood surfaces and mixing times of MCMC methods for phylogenetic inference.
2. The local-to-global property in phylogenetics[4].
3. ℓ_1-type regularized estimator to infer sampled ancestors.

[2] meaning that it can reconstruct all edges of length greater than any given threshold from gene sequences of polynomial length.

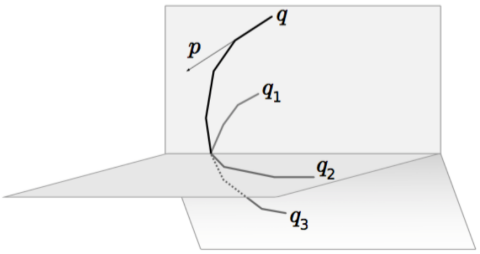

Figure 3: The phylogenetic likelihood function is non-differentiable across topologies and the tree space is not a manifold. phyloHMC traverses the tree space using a stochastic extension of the standard leap-frog integrator.

[3] namely, time-reversibility, volume preservation and accessibility

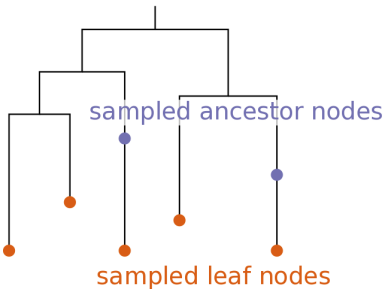

Figure 4: Unlike standard setting where all data are sampled at leaves, ℓ_1-type regularized estimator allows sampled data to belong to an ancestor nodes.

[4] namely, how information from a single tree topology can be used to represent the likelihood from on the whole tree space

Probabilistic methods for experimental design and control of dynamical systems

Studies of biological systems are usually hindered by several factors: biological systems are usually unidentifiable, and data collected to study such systems are often very sparse and noisy due to technical limitations and experimental constraints. Moreover, the presence of multivariate bifurcations often leads to system behaviors that are very different in nature. As a result, analyses of high-dimensional biological systems usually need to be performed locally.

In my work, we resolve those issues by building a probabilistic framework for experimental design and control of biological systems and utilizing this framework to obtain effective sampling schemes for global systems analysis. We develop a sequential scheme for experimental design and prove its consistency [5]. In a similar manner, we analyze different effective sampling schemes for behavior discrimination and provide theoretical support for such approaches [6]. We also employ this framework and the behavior discrimination algorithm to derive explicit model predictive controllers of the T-cell signaling pathway by model-informed open-loop control [7] and integral sliding mode control (Figure 5) [8, 9].

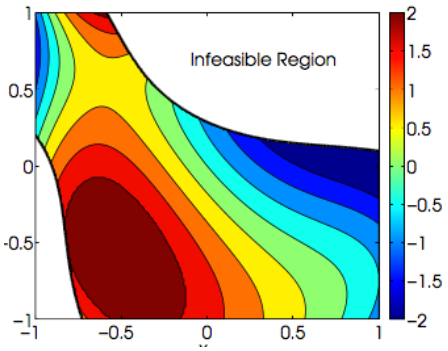

Figure 5: A model explicit controller constructed using a behavior discrimination algorithm with low-descrepancy sampled data.

Future directions:

1. Data-free global identifiability of biological systems.
2. Uncertainty quantification methods for models with discontinuous responses.

Machine learning: theory and applications

The rate with which a learning algorithm converges as more data come in plays a central role in machine learning. I am interested in settings under which fast learning rates[5] are possible.

Learning with non-regular data. In various applications, the standard assumption for statistical learning, which dictates that data are distributed independent and identically corresponding to a well-behaved distribution, may not hold. This makes the tasks of designing and analyzing learning algorithms more challenging. We attempt to relax such conditions in various settings. First, we derive the convergence rate of the *weighted average algorithm* when the training data is a V-geometrically ergodic Markov chain [10]. We then prove new fast learning rates for *one-vs-all multi-class plug-in classifiers* trained from mixing data [11]. In [12], we obtain fast learning rate for the *empirical risk minimization estimator* when the distribution of the losses over the hypothesis spaces has heavy tails.

Applications. In addition to theory, I am also interested in applications of machine learning algorithms in applied sciences. In [13], we develop a spectral-based representation method with more than 90% accuracy in identifying individuals from their eye movements. We also employ *active learning*[6] schemes to increase performances of learning algorithms in various applications, including systems biology [5, 6], control theory [8, 9] and applied spectroscopy [14].

Future directions:

1. Pseudo-Bayesian learning with heavy-tailed losses.
2. Bayesian pool-based active learning with weak labelers[7].
3. Designing optimal sampling schemes for phylogenetic models.

[5] "Fast learning" means the trained classifier converges with rate faster than $n^{-1/2}$.

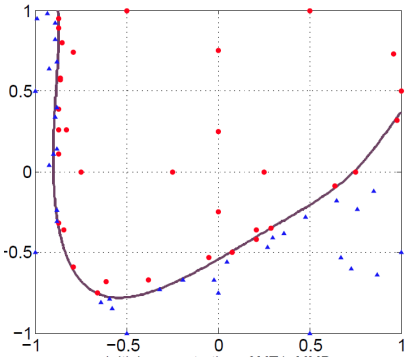

Figure 6: Sampled data for classification from a pseudo-Bayesian active sampling scheme.

[6] Active learning means the learning algorithm is able to interactively query labelers to maximize the information received and increase algorithmic performance.

[7] In contrast to standard settings, weak labelers have the option to abstain from providing the labels of an instance.

References

[1] Vu Dinh and Frederick A Matsen IV. The shape of the one-dimensional phylogenetic likelihood function. *The Annals of Applied Probability*, 2016.

[2] Vu Dinh, Lam Si Tung Ho, Marc A Suchard, and Frederick A Matsen IV. Consistency and convergence rate of phylogenetic inference via regularization. *arXiv preprint arXiv:1606.03059*, 2016.

[3] Vu Dinh and Frederick A Matsen IV. Hamiltonian Monte Carlo on the orthant complex of phylogenetic trees. *In preparation*, 2016.

[4] Vu Dinh, Aaron E Darling, and Frederick A Matsen IV. Online Bayesian phylogenetic inference: theoretical foundations via Sequential Monte Carlo. *In review*, 2016.

[5] Vu Dinh, Ann E Rundell, and Gregery T Buzzard. Experimental design for dynamics identification of cellular processes. *Bulletin of Mathematical Biology*, 76(3):597–626, 2014.

[6] Vu Dinh, Ann E Rundell, and Gregery T Buzzard. Effective sampling schemes for behavior discrimination in nonlinear systems. *International Journal for Uncertainty Quantification*, 4(6), 2014.

[7] Jeffrey P Perley, Judith Mikolajczak, Vu Dinh, Marietta L Harrison, Gregery T Buzzard, and Ann E Rundell. Systematically manipulating T-cell signaling dynamics via multiple model informed open-loop controller design. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 380–385.

[8] Ankush Chakrabarty, Vu Dinh, Gregery T Buzzard, Stanislaw H Żak, and Ann E Rundell. Robust explicit nonlinear model predictive control with integral sliding mode. In *2014 American Control Conference (ACC)*, pages 2851–2856.

[9] Ankush Chakrabarty, Vu Dinh, Martin Corless, Ann E Rundell, Stanislaw H Zak, and Gregery T Buzzard. SVM-informed explicit nonlinear model predictive control using low-discrepancy sequences. *IEEE Transactions on Automatic Control*, 2016.

[10] Nguyen Viet Cuong, Lam Si Tung Ho, and Vu Dinh. Generalization and robustness of batched weighted average algorithm with V-geometrically ergodic Markov data. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 264–278. Springer, 2013.

[11] Vu Dinh, Lam Si Tung Ho, Nguyen Viet Cuong, Duy Nguyen, and Binh T Nguyen. Learning from non-iid data: Fast rates for the one-vs-all multiclass plug-in classifiers. In *International Conference on Theory and Applications of Models of Computation (TAMC)*, pages 375–387. Springer, 2015.

[12] Vu Dinh, Lam Si Tung Ho, Duy Nguyen, and Binh T Nguyen. Fast learning rates with heavy-tailed losses. In *Conference on Neural Information Processing Systems (NIPS)*, 2016.

[13] Nguyen Viet Cuong, Vu Dinh, and Lam Si Tung Ho. Mel-frequency cepstral coefficients for eye movement identification. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 253–260, 2012.

[14] Owen G. Rehrauer, Vu Dinh, Bharat Mankani, Gregey T. Buzzard, Bradley Lucier, and Dor Ben-Amortz. Binary-complementary compressive filters for Raman spectroscopy. *In preparation*, 2016.