Joint maximum-likelihood of phylogenies and ancestral states is not consistent

David A. Shaw[1] and Frederick A. Matsen IV[*1]

[1]Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Abstract

Maximum likelihood estimation in phylogenetics requires a means of handling unknown ancestral states. Classical maximum likelihood averages over these unknown intermediate states, leading to consistent estimation of the topology and continuous model parameters. Recently, a computationally-efficient approach was proposed to jointly maximize over these unknown states and phylogenetic parameters. Although this method of joint maximum likelihood estimation can obtain estimates more quickly, its properties as an estimator are not yet clear. In this paper, we show that this method of jointly estimating phylogenetic parameters along with ancestral states is not consistent in general. We find a set of parameters that generate data under a four-taxon tree for which this joint method estimates an incorrect topology in the limit of infinite-length sequences. For branch length estimation on the correct topology, we outline similar cases where branch length estimates are consistently and heavily biased.

[*]Corresponding author, email: matsen@fredhutch.org

Introduction

Classical maximum likelihood (ML) estimation in phylogenetics operates by integrating out latent ancestral states at the internal nodes of the tree. In a recent paper, Sagulenko et al. [2017] suggest using an approximation to ML inference in which the likelihood is maximized jointly across model parameters and ancestral sequences on a fixed topology. This is attractive from a computational perspective: such joint inference can proceed according to an iterative procedure in which ancestral sequences are first estimated and model parameters are optimized conditional on these estimates. This latter conditional optimization is simpler and more computationally efficient than optimizing the marginal likelihood. But is it statistically consistent?

An estimator is said to be statistically consistent if it converges to the generating model with probability 1 in the large-data limit; existing consistency proofs for maximum likelihood phylogenetics [RoyChoudhury et al., 2015] apply only to estimating model parameters when the ancestral sequences have been integrated out of the likelihood. These proofs do not readily extend to include estimating ancestral states. Moreover, examples of inconsistency arising from problems where the number of parameters increases with the amount of data [Neyman and Scott, 1948] indicate that joint inference of trees and ancestral states may not enjoy good statistical properties. In this case those additional parameters come in the form of the states of ancestral sequences. Although the software described in Sagulenko et al. [2017] fits on a user-supplied topology and the authors explicitly warn that the approximation is for the case where "branch lengths are short and only a minority of sites change on a given branch," their work motivates understanding the general properties of such joint inference. In particular, one would like to know when this approximate technique breaks down for both topology and branch length inference, even when sequence data is "perfect," i.e., is generated without sampling error according to the exact model used for inference.

In this paper, we show that the joint inference of trees and ancestral se-

quences is not consistent in general. To do so, we use a binary symmetric model with data being generated on the four-taxon "Farris zone" [Siddall, 1998] tree, and we construct bounds on the joint objective function to demarcate a sizeable area of long branch lengths in which joint inference is guaranteed to give the wrong tree in the case of perfect sequence data with an infinite number of sites. We find similar areas where joint inference consistently overestimates interior branch lengths when the topology is known and fixed.

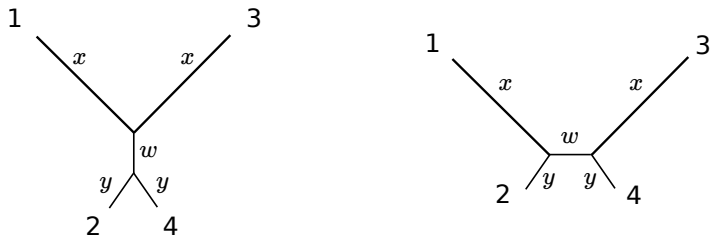
Phylogenetic maximum likelihood

Assume the binary symmetric model, namely with a character alphabet $\mathcal{A} = \{0, 1\}$ and a uniform stationary distribution [Semple and Steel, 2003]. Let m be the number of tips of the tree, and $p = m-2$ the number of internal nodes. We observe n independent and identically distributed samples of character data, i.e., an alignment with n columns, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathcal{A}^{m \times n}$ distributed as the random variable Y. The corresponding unobserved ancestral states are $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n] \in \mathcal{A}^{p \times n}$ and distributed as H.

We parameterize branches on the unique unrooted four-tip phylogenetic tree in ways known as the "Farris" and "Felsenstein" trees (Fig. 1). In the standard configuration of each of these trees, the interior branch length parameters are equal to the bottom two parameters. We show in the Appendix that, in the case of the Farris tree, performing inference fixing both the top two branch parameters to be equal and the bottom two branch parameters to be equal will obtain the same maximum likelihood estimate as in the case of arbitrary branch parameters.

We parameterize the branches of these trees not with the standard notion of branch length in terms of number of substitutions per site, but with an alternate formulation called "fidelity." The probability of a substitution on a branch with fidelity θ is $(1-\theta)/2$ while the probability of no substitution is $(1+\theta)/2$ where $0 \leq \theta \leq 1$. This parameter quantifies the fidelity of transmission of the ancestral state across an edge [Matsen and Steel, 2007].

Fidelities have useful algebraic properties, and generating probabilities



(a) Farris tree τ_1 (b) Felsenstein tree τ_2

Figure 1: Two four-taxon trees with fidelities as labeled: $\theta_1 = \theta_3 = x$, $\theta_2 = \theta_4 = y$, and $\theta_5 = w$. {fig:farris-fels-top

using the Hadamard transform have an especially simple form (see (8) in the Appendix). For a four-taxon tree, define the general branch fidelity parameter $t = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$ where fidelities are ordered in the order of the taxa with the internal branch last (Fig. 1).

Two paths to maximum likelihood

The standard phylogenetic likelihood approach on unrooted trees under the usual assumption of independence between sites is as follows. For a topology τ and branch fidelities t the likelihood given observed ancestral states $\mathbf{H}$ is

$$L_n(\tau, t; \mathbf{Y}, \mathbf{H}) = \prod_{i=1}^{n} \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t). \quad (1) \quad \{\texttt{eq:full\_likelihood}\}$$

The probability $\Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t)$ is a product of transition probabilities determined by $\mathbf{Y}$, $\mathbf{H}$, τ, and t [Felsenstein, 2004].

The classical approach is to maximize the likelihood marginalized across ancestral states

$$\tilde{L}_n(\tau, t; \mathbf{Y}) = \prod_{i=1}^{n} \sum_{\mathbf{h}_i \in \mathcal{A}^p} \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) \quad (2) \quad \{\texttt{eq:marginal\_likelih}$$

to estimate the tree τ and branch fidelities t.

The alternative approach [Sagulenko et al., 2017] does away with the marginalization and directly estimates the maximum likelihood parameters of the fully-observed likelihood in (1). This is known in statistics as a profile likelihood [Murphy and van der Vaart, 2000], which exists here because $\mathcal{A}$ is a finite set:

$$L'_n(\tau, t; \mathbf{Y}) = \prod_{i=1}^{n} \max_{\mathbf{h}_i \in \mathcal{A}^p} \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) = \max_{\mathbf{H} \in \mathcal{A}^{p \times n}} L_n(\tau, t; \mathbf{Y}, \mathbf{H}). \tag{3}$$ {eq:profile_likelihc

We use $\hat{\mathbf{H}}$ to denote an estimate for $\mathbf{H}$ obtained by maximizing (3), and estimate a topology and branch fidelities using this profile likelihood as

$$(\hat{\tau}, \hat{t}) = \operatorname*{argmax}_{\tau, t} L'_n(\tau, t; \mathbf{Y}). \tag{4}$$ {eq:profile_likelihc

In general, the functional form of (3) is determined by inequalities that depend on the unknown (τ, t). For this reason, in practice, the joint inference strategy estimates $\hat{\mathbf{H}}$ for a fixed (τ, t), then $(\hat{\tau}, \hat{t})$ given $\hat{\mathbf{H}}$, maximizing each of these conditional objectives until convergence [Sagulenko et al., 2017].

Inconsistency of joint inference

We now state our results on the inconsistency of joint inference. All proofs are deferred to the Appendix.

Assume $\mathbf{Y}$ is generated from topology τ^* and branch fidelities t^*. Use $\ell_{\tau^*, t^*}(\tau, t)$ to denote the expected per-site log-likelihood, which can be thought of as the infinite-length sequence case

$$\frac{1}{n} \log L'_n(\tau, t; \mathbf{Y}) \to \ell_{\tau^*, t^*}(\tau, t).$$

We give ℓ explicitly as (6) in the Appendix.

Inconsistency in topology estimation

To show an inconsistency in topology estimation, we start with true generating parameters $t^* = \{x^*, y^*, x^*, y^*, y^*\}$ on the Farris topology (Fig. 1a). We show that, as $n \to \infty$, there exist values for x^* and y^* such that the value of the likelihood after maximizing using joint inference is greater for the Felsenstein topology than for the true, generating Farris topology. To do so, we construct an upper bound $C_0(x^*, y^*)$ for the likelihood given the Farris topology as a function of x^* and y^* and, similarly, a lower bound $C_1(x^*, y^*)$ for the likelihood given the Felsenstein topology. When $C_0(x^*, y^*) < C_1(x^*, y^*)$, the likelihood in the Felsenstein case is larger than the likelihood in the Farris case, demonstrating inconsistency (Fig. 2).

{thmt@@topoInconsist
{thmt@@topoInconsist

Theorem 1. *Let* $t^* = \{x^*, y^*, x^*, y^*, y^*\}$ *and* $t = \{x, y, x, y, w\}$. *There exist* $C_0(x^*, y^*), C_1(x^*, y^*)$, *and a set of* $0 < x^*, y^* < 1$ *such that*

$$\max_t \ \ell_{\tau_1, t^*}(\tau_1, t) \leq C_0(x^*, y^*),$$

$$C_1(x^*, y^*) \leq \max_t \ \ell_{\tau_1, t^*}(\tau_2, t)$$

with $C_0(x^*, y^*) < C_1(x^*, y^*)$.

The proof of this theorem is by a detailed examination of inequalities. Intuitively, τ_2 is favored in performing joint inference since the objective function for τ_2 has more "degrees of freedom"—Table S4 shows that τ_1 has only three possible forms for its objective function while τ_2 has many more. This enables more possible maxima for τ_2 even if data are generated from τ_1.

Inconsistency in branch length estimation

We now consider the problem of branch length estimation on the correct tree using joint estimation. As described above, we use the equivalent but different notion of branch fidelities. We analyze two settings on the Farris tree, corresponding to whether some branch fidelities are fixed at their true values or not. As above, assume that data is generated from the Farris tree

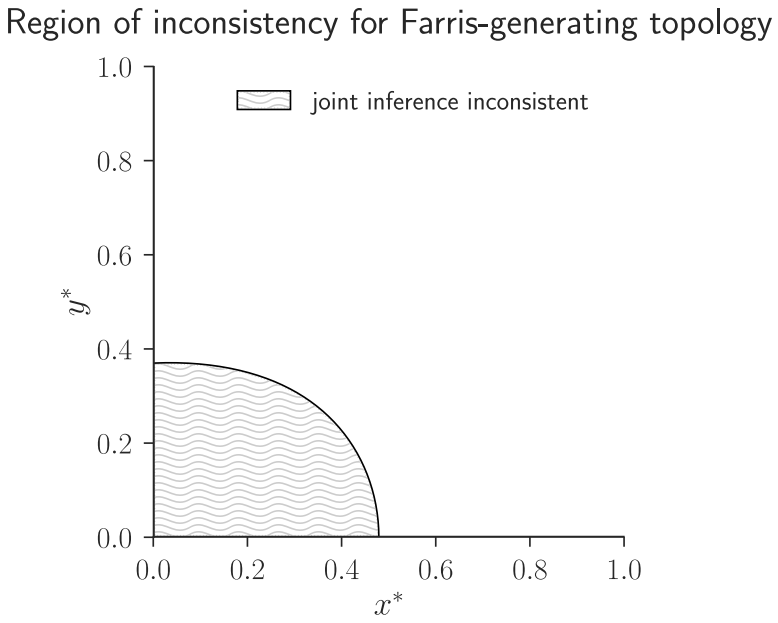


Figure 2: An analytically-derived region of topological inconsistency in terms of fidelities for "perfect" data generated on the Farris topology (Fig. 1) with $w^* = y^*$. Due to the looseness of the upper and lower bounds, the parameters in the white region do not necessarily indicate consistency, though all parameters in the shaded region result in an inconsistency. {fig:inconsistency-f

with two top branches of fidelity x^* and all other branches of fidelity y^* (Fig. 1). In the first "restricted" case, we show that for a nontrivial subset of possible values for x^* and y^*, the interior branch fidelity parameter w will be consistently overestimated as exactly equal to one (zero branch length) instead of its true value of y^*. That is, if we estimate x and y correctly, then, for

$$\hat{w} = \arg\max_{w} \ \ell_{\tau_1,t^*}(\tau_1, \{x^*, y^*, x^*, y^*, w\}),$$

there is a set of values for x^* and y^* where $\hat{w} \equiv 1$. In the general case, we do the same but with

$$(\hat{x}, \hat{y}, \hat{w}) = \arg\max_{x,y,w} \ \ell_{\tau_1,t^*}(\tau_1, \{x, y, x, y, w\})$$

to find a region where the inferred values do not converge to the generating values. These situations are in contrast to the approach using marginal likelihood where $\hat{w}$ necessarily converges to y^* as the number of observations grows [RoyChoudhury et al., 2015].

Restricted case

Fix estimated fidelities $x = x^*$ and $y = y^*$ to their true, generating values and estimate the internal branch parameter w.

{thmt@@restrictedBra
{thmt@@restrictedBra

Theorem 2. *Let*

$$\beta := \beta(x^*, y^*) = 1 + (x^*)^2 + (y^*)^2 + (x^*)^2(y^*)^2,$$

$$\gamma := \gamma(x^*, y^*) = 4x^*y^*.$$

The maximum likelihood value $\hat{w}$ is equal to 1 if

$$-\gamma^2\left(1 + \frac{1}{2}\beta\right) + 2\gamma\beta x^*(y^*)^2 + \beta^2 \geq 0$$

and there exists a set of $0 < x^*, y^* < 1$ satisfying this.

This theorem allows us to demarcate a region of biased internal branch

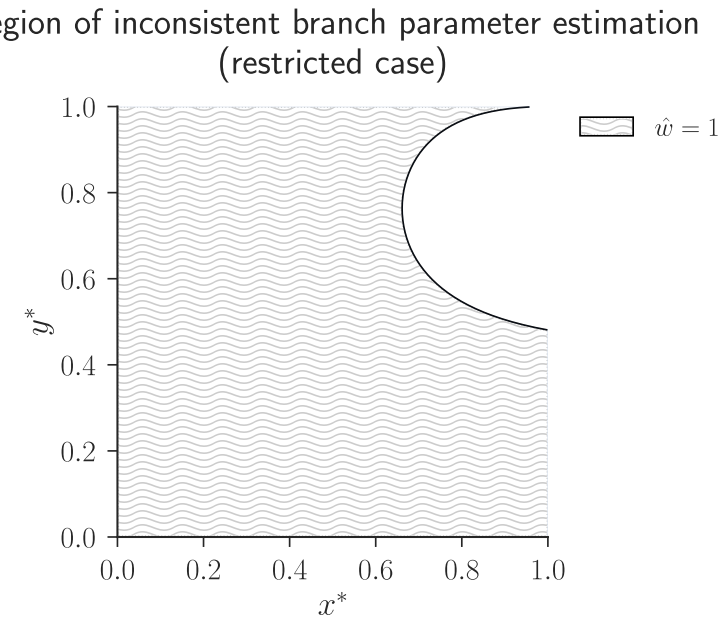


Figure 3: Analytically-derived region of branch parameter inconsistency in terms of fidelities $x = x^*$ and $y = y^*$ that are fixed to their correct values with the same data-generating setup as Fig. 2. The shaded region shows the area in which the internal branch length is estimated to be 0, i.e. the estimated fidelity $\hat{w} = 1$, even though the generating fidelity is y^*. Here again, the shaded region is guaranteed to give inconsistent estimation, while the white region may or may not do so.

{fig:bl-inconsistenc

length estimation by plotting where the inequality in Theorem 2 is satisfied (Fig. 3). Intuitively, this happens when estimated ancestral sequences at internal nodes are identical across a branch with this branch length estimated to be zero (i.e., has fidelity $\hat{w} = 1$). As an intuition for the theoretical development, seeing no change along a branch more likely increases the likelihood by introducing a term of $(1 + \theta)$ instead of $(1 - \theta)$, and branch fidelities will be positively biased due to this. If we allow multifurcating trees in our inference, then we can think of this as another instance of converging to the wrong topology.

General case

The general case is more challenging to analyze and so we obtain weaker bounds. Here, $\hat{w}$ is a function of x^*, y^*, $\hat{x}$, and $\hat{y}$. Looking to the previous section, the region where $\hat{w} = 1$ will still be given by the inequality in Theorem 2, only with γ and β now being functions of $\hat{x}$ and $\hat{y}$ instead of x^* and y^*. Assume we know $\hat{x}$ and $\hat{y}$ as functions of x^* and y^*. We show there are similar bounds as in the restricted case, though we need to take into account the unknown values of $\hat{x}$ and $\hat{y}$. We fix bounds on these estimates and show that, in the general case, joint estimation either estimates $\hat{w}$ to be one or estimates $\hat{x}$ or $\hat{y}$ to fall outside of specified bounds, indicating a poor estimate in at least one of the three unknown branch parameters.

{thmt@@generalBranch
{thmt@@generalBranch

Theorem 3. *Define* $\gamma(x, y) = 4xy$. *For*

$$\beta := \beta(x^*, y^*) = 1 + (x^*)^2 + (y^*)^2 + (x^*)^2(y^*)^2,$$

$$\gamma := \gamma(x^*, y^*),$$

bounds

$$\gamma_L := \gamma_L(x^*, y^*) \leq \gamma(\hat{x}, \hat{y}),$$

$$\gamma_U := \gamma_U(x^*, y^*) \geq \gamma(\hat{x}, \hat{y}),$$

and

$$\beta_L := \beta_L(x^*, y^*) \leq \beta(\hat{x}, \hat{y}),$$

the maximum likelihood value $\hat{w} = 1$ *when*

$$-\gamma_U^2 \left(1 + \frac{1}{2}\beta\right) + 2\gamma_L\beta_L x^*(y^*)^2 + \beta_L^2 \geq 0.$$

We use this theorem to show incorrect branch parameter estimates as follows. If we do not tolerate any error in pendant branches, we use the tightest possible bounds $\gamma_L = \gamma_U = \gamma$ and $\beta_L = \beta$, which is the restricted case of the previous section (Fig. 3). For an intermediate bound, define a

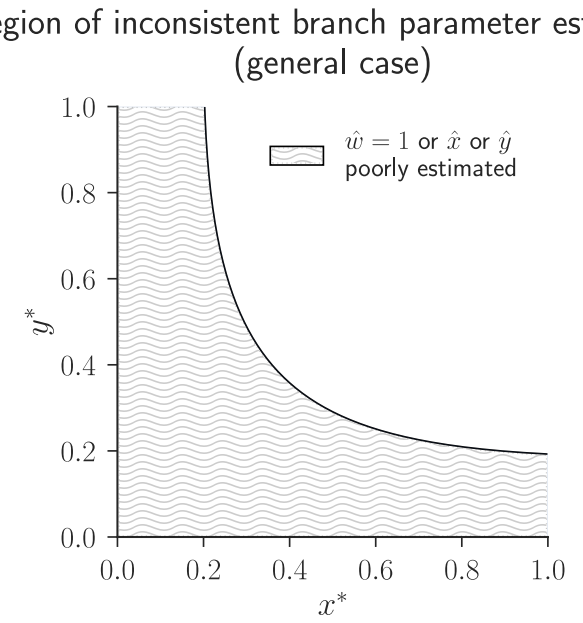


Figure 4: Analytically-derived region of branch parameter inconsistency in terms of fidelities with the same data-generating setup as Fig. 2. In the marked region, either $\hat{w} = 1$ or one of $x^* - 0.1 \leq \hat{x} \leq x^* + 0.1$ or $y^* - 0.1 \leq \hat{y} \leq y^* + 0.1$ will not be true, resulting in poor estimation of the pendant branch parameters. As in previous plots, the shaded and white regions are loose indications of inconsistency.

{fig:bl-loose-incons

specified allowable level of error in estimates for $\hat{x}$ and $\hat{y}$ so that

$$x^* - x_L \leq \hat{x} \leq x^* + x_U$$

and

$$y^* - y_L \leq \hat{y} \leq y^* + y_U.$$

Then γ_L, γ_U and β_L can be derived directly from the bounds on $\hat{x}$ and $\hat{y}$. As an example, the region to the left side of the curve in Fig. 4 shows the case where $x_L = x_U = y_L = y_U = 0.1$. In this case, joint inference will either estimate $\hat{w}$ to have fidelity one or estimate either $\hat{x}$ or $\hat{y}$ to be more than 0.1 away from its true fidelity.

Empirical validation

Direct numerical optimization confirms our theoretically-derived bounds and provides a more detailed picture compared to the conservative analytically-derived region (Fig. 4). To determine how conservative, we use the method of basin-hopping [Wales and Doye, 1997] to perform joint estimation (Fig. 5). We see that the region of inconsistency in the general case is similar to that of the restricted case (compare Figs. 3 and 5). This region encompasses the majority of the branch fidelity space; even given the correct topology and performing our best possible optimization, we have many situations where we will estimate the interior branch fidelity to be one.

We provide a full description of our optimization procedure in the Appendix, but briefly, we perform two maximizations—one over $0 \leq x, y, w \leq 1$ and one over $0 \leq x, y \leq 1$ with $w = 1$—and take the value of $\hat{w}$ with the higher objective function. We compute these maxima over a lattice in steps of 10^{-2} for x^* and y^* from 10^{-2} to $1 - 10^{-2}$. We do not include zero or one in our lattice to further stabilize the fits, as these cases can result in pathologies. Our optimization code can be found at `https://github.com/matsengrp/joint-inf/`.

Marginal inference performs as expected, where $\hat{w}$ is equal to y^* regardless of the value of x^* (Fig. S2) when optimizing (2) using the same procedure. For joint inference, the estimates for $\hat{w}$ when x^* and y^* are both large look reasonable, with $\hat{w}$ increasing as y^* increases, though Fig. S3 shows there is a systematic positive bias in this procedure even when $\hat{w}$ is not estimated to be one. To understand the quality of each fit, we report the range of $\hat{w} - y^*$ where $\hat{w} \neq 1$. For joint inference, the errors range from $[-7 \times 10^{-3}, 8 \times 10^{-2}]$ and for marginal inference, $[-8 \times 10^{-8}, 5 \times 10^{-7}]$ showing that, even in cases where joint inference does not estimate $\hat{w}$ to be exactly one, it still fails to achieve a low error from truth.

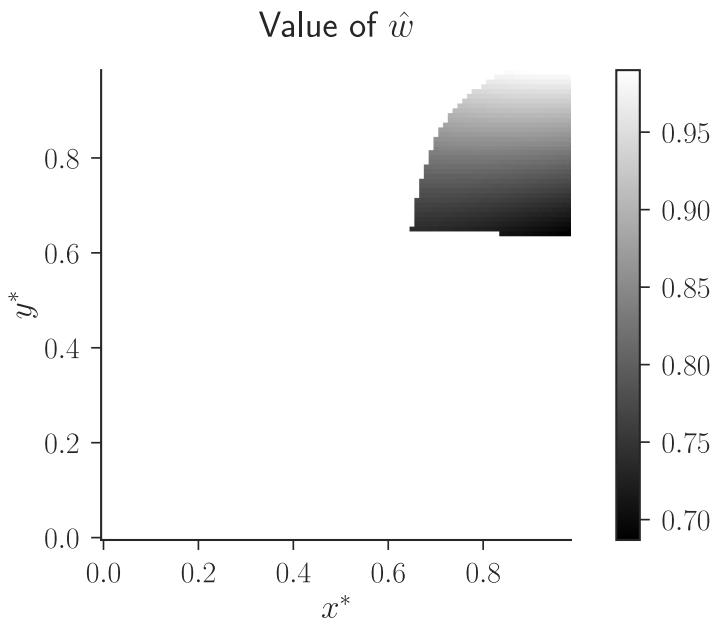


Figure 5: Numerical estimates for $\hat{w}$ when computing $(\hat{x}, \hat{y}, \hat{w})$ using basin-hopping [Wales and Doye, 1997] optimizing (3). Data generated as in Fig. 2. {fig:bl-general-incc

Discussion

We have shown that jointly inferring ancestral states and phylogenetic parameters [Sagulenko et al., 2017] is not consistent in general. Specifically, in the case of four-taxon trees with infinite data, we have obtained nontrivial regions of generating parameters that result in two types of inconsistency: first, where joint inference converges on the incorrect topology and, second, where it estimates severely biased branch lengths even given the correct topology. In all cases, these regions of inconsistency arise when the branches of the generating trees are "long," that is, when branch fidelities tend to be small. This inconsistency in the case of long branches concurs with some empirical findings in Sagulenko et al. [2017], namely their Figures 2 and 3.

Joint inference of tree parameters and ancestral sequences is a type of profile likelihood, a well-studied subject in statistics [Murphy and van der Vaart, 2000]. Many properties regarding the performance of maximum likelihood estimates obtained using this approach are known, and many methods exist to overcome their undesirable properties, e.g., the method of sieves [Geman and Hwang, 1982]. A potential solution in this case using the method of sieves could be to project the column-wise ancestral state patterns into a lower-dimensional space, allowing the degrees of freedom in the ancestral state columns to grow with n, albeit more slowly than $O(n)$. Elsewhere in statistics literature, the failure of maximum likelihood estimates to obtain consistent estimates as the number of parameters goes to infinity have been shown by the Neyman-Scott paradox [Neyman and Scott, 1948], though parameters tending to infinity is not a necessary condition for inconsistency [Le Cam, 1990]. Consistency proofs of standard maximum likelihood estimates of phylogeny (2) are recent [RoyChoudhury et al., 2015], and no results have been obtained for profile likelihood. We have furthered progress in understanding the limitations of this joint optimization procedure.

Previous work in phylogenetics has developed consistency counterexamples using the same four-taxon topologies used here [Felsenstein, 1978].

In this previous work, when simulating data under the "Felsenstein zone" topology τ_2, as the number of observations increases, the "Farris zone" topology τ_1 becomes more likely when performing a particular estimation procedure. This is the converse of what we have shown for joint inference, where the Felsenstein topology is more likely than the Farris topology. Moreover, the inconsistency demonstrated by Felsenstein [1978] is attributed to long branch attraction, i.e., the fact that there may be multiple long branches where parallel changes are more likely than a single change along a short branch. This is not the case here; for our case, the inconsistency generally occurs when all branches are long, and has more to do with the choices of the form of the likelihood than from the interplay of long and short branches. Difficulties in phylogenetic estimation when generating data on the "Farris zone" tree have been found by Siddall [1998], though Swofford et al. [2001] show that sequence length plays a major role in these issues.

While we have shown inconsistency in both topology and branch parameter estimation, there is substantial scope for future work to make these results more precise and more general. The techniques used to obtain upper and lower bounds for the likelihoods in the topology estimation case provide relatively loose bounds, though how loose they are remain unknown without either further analysis or verification through simulation. Similarly, for the general case in estimating branch lengths, we were only able to provide a conservative region of overestimation, and the unusual shape we observe via numerical optimization (Fig. 5) begs further investigation. Empirical validation shows that the general case is not unlike the restricted case. All of these results hold only for a simple binary symmetric model on four-taxon trees, and extensive simulation is necessary to understand how these results extend to more complicated general cases. Given that many of the bounds presented here are in the form of level sets of multivariate polynomials, a more formal approach using algebraic geometric techniques may reveal more stable or interesting patterns of inconsistency; see Sturmfels [2002] for a thorough treatment of solving systems of polynomial equations. Finally, all of the material presented here concerns joint

estimation under maximum likelihood, and does not pose any problem for other settings, such as joint sampling of trees and ancestral sequences in a Bayesian framework.

Acknowledgements

We thank Richard Neher, Vladimir Minin, and Joe Felsenstein for helpful discussions.

This work was supported by National Institutes of Health grants R01-AI12096, U19-AI117891, and U54-GM111274 as well as National Science Foundation grants CISE-1561334 and CISE-1564137. The research of Frederick Matsen was supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute and the Simons Foundation.

References

Joseph Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410, 1 December 1978. ISSN 0039-7989. doi: 10.2307/2412923. URL http://www.jstor.org/stable/2412923.

Joseph Felsenstein. *Inferring phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts, 2004.

Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–414, 1982.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/. [Online; accessed 20 Nov 2017].

Lucien Le Cam. Maximum likelihood: An introduction. *International Statistical Review*, 58(2):153–171, Aug 1990.

Frederick A. Matsen and Mike Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, 56(5):767–775, 1 October 2007. ISSN 1063-5157. doi: 10.1080/10635150701627304. URL http://sysbio.oxfordjournals.org/content/56/5/767.abstract.

Susan A. Murphy and Aad W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000. ISSN 0162-1459. doi: 10.2307/2669386. URL http://www.jstor.org/stable/2669386.

Jerzy Neyman and Elizabeth L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948. ISSN 0012-9682, 1468-0262. doi: 10.2307/1914288. URL http://www.jstor.org/stable/1914288.

Arindam RoyChoudhury, Amy Willis, and John Bunge. Consistency of a phylogenetic tree maximum likelihood estimator. *Journal of Statistical Planning and Inference*, 161:73–80, June 2015. ISSN 0378-3758. doi: 10.1016/j.jspi.2015.01.001. URL http://www.sciencedirect.com/science/article/pii/S0378375815000038.

Pavel Sagulenko, Vadim Puller, and Richard Neher. TreeTime: maximum likelihood phylodynamic analysis. 21 June 2017. URL http://biorxiv.org/content/early/2017/06/21/153494.

Charles Semple and Mike Steel. *Phylogenetics*. Oxford University Press, New York, NY, 2003.

Mark E Siddall. Success of parsimony in the Four-Taxon case: Long-Branch repulsion by likelihood in the farris zone. *Cladistics*, 14(3):209–220, 1 September 1998. ISSN 0748-3007, 1096-0031. doi: 10.1111/j.1096-0031.1998.tb00334.x. URL http://dx.doi.org/10.1111/j.1096-0031.1998.tb00334.x.

Bernd Sturmfels. Solving systems of polynomial equations. In *American Mathematical Society, CBMS Regional Conferences Series, No. 97*, 2002.

David L. Swofford, Peter J. Waddell, John P. Huelsenbeck, Peter G. Foster, Paul O. Lewis, and James S. Rogers. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology*, 50(4):525–539, August 2001. ISSN 1063-5157. URL http://www.ncbi.nlm.nih.gov/pubmed/12116651.

David J. Wales and Jonathan P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.

Appendix

Site split formulation

We begin by introducing "site splits," which formalize the notion that a given site pattern is equally probable to its complement under the binary symmetric model. This is a standard step in the description of the Hadamard transform (Section 8.6 of Semple and Steel [2003]), although our approach is complicated slightly by the inclusion of ancestral states.

Since we have a finite character alphabet, for a given column i there are a finite number of possible assignments of characters to tips $\mathbf{y}_i$ or internal nodes $\mathbf{h}_i$; this results in a simplification of likelihood calculation. Take the tip labels of τ to be $\{1, \ldots, m\}$. For likelihood calculation under the binary symmetric model, we describe a given $\mathbf{y}_i$ as a subset of indices $\tilde{y} \subseteq \mathcal{Y} := \{1, \ldots, m-1\}$ with equivalent characters, commonly called a "site split." We define the site split $\tilde{y}$ for a $\mathbf{y}_i$ as simply $\mathbf{y}_i$ if the label m is not in $\mathbf{y}_i$, and as its complement otherwise. Taking such a complement simplifies but does not change the result of likelihood computation because the probability of observing a particular collection of binary characters is equivalent to the probability of its complement under the binary symmetric model.

For topology τ, we define an ordered set of internal node labels $\{1, \ldots, p\}$ for $\mathbf{h}_i$ and similarly use a subset of characters $\tilde{h} \subseteq \mathcal{H} := \{1, \ldots, p\}$ to describe a realization $\mathbf{h}_i$. In this case the entire set of internal nodes must be enumerated: the probability of observing an ancestral state split conditional on a site split is not invariant to taking its complement.

We enumerate the site splits $\tilde{y}_j$ of which there are $q = |\mathcal{P}(\mathcal{Y})|$ in total where $\mathcal{P}$ denotes the power set. Similarly we enumerate ancestral splits $\tilde{h}_k$ of which there are $r = |\mathcal{P}(\mathcal{H})|$ in total.

We first fix notation.

Definition. *Let the mapping from site patterns to site splits be*

$$\psi : \mathcal{A}^m \to \mathcal{P}(\mathcal{Y})$$

and the mapping from ancestral states and tip states to ancestral state splits be

$$\xi : \mathcal{A}^p \times \mathcal{A}^m \to \mathcal{P}(\mathcal{H}).$$

Then, given a site pattern–valued random variable Y, define the random variable

$$\Psi := \psi(Y)$$

that takes corresponding realizations $\tilde{y}_j$ for some j, and

$$\Xi := \xi(Y, H)$$

for a tip state–valued random variable Y and an ancestral state–valued random variable H.

The mapping ψ takes the complement of site patterns to obtain a site split in $\mathcal{P}(\mathcal{Y})$. The mapping ξ is defined by whether the tip states have their complements taken or not: if a set of tip labels $\mathbf{y}$ is in $\mathcal{Y}$, $\xi(\mathbf{y}, \mathbf{h})$ is $\mathbf{h}$; otherwise, if $\mathbf{y}$ is not in $\mathcal{Y}$, then the complement of $\mathbf{y}$ necessarily is in $\mathcal{Y}$, and $\xi(\mathbf{y}, \mathbf{h})$ is the complement of $\mathbf{h}$.

For the ith factor of (1),

$$\Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) = \Pr(Y = \mathbf{y}_i \mid \tau, t) \cdot \Pr(H = \mathbf{h}_i \mid Y = \mathbf{y}_i, \tau, t).$$

As a consequence of assuming a binary symmetric model, taking complements yields

$$\begin{aligned} 2 \cdot \Pr(Y = \mathbf{y}_i \mid \tau, t) &= \Pr(\Psi = \psi(\mathbf{y}_i) \mid \tau, t) \\ &= \Pr(\Psi = \tilde{y}_j \mid \tau, t) \end{aligned}$$

for some j and

$$\Pr(H = \mathbf{h}_i \mid Y = \mathbf{y}_i, \tau, t) = \Pr(\Xi = \xi(\mathbf{y}_i, \mathbf{h}_i) \mid \Psi = \psi(\mathbf{y}_i), \tau, t).$$

Given (τ, t), there exists an ordered list of sets $\boldsymbol{\eta}(\tau, t) = (\eta_1(\tau, t), \ldots, \eta_q(\tau, t))$

such that any element ξ_j of the jth component $\eta_j(\tau, t)$ satisfies

$$\max_{\tilde{h}_k \in \mathcal{P}(\mathcal{H})} \Pr(\Xi = \tilde{h}_k \mid \Psi = \tilde{y}_j, \tau, t) = \Pr(\Xi = \xi_j \mid \Psi = \tilde{y}_j, \tau, t).$$

In other words, for the jth site split, $\eta_j(\tau, t) \subset \mathcal{P}(\mathcal{H})$ is the set of most likely ancestral splits for that particular site split, topology and set of branch lengths, and ξ_j is one of possibly many equiprobable ancestral state splits in $\eta_j(\tau, t)$. For each $\mathbf{y}_i$, $\xi(\mathbf{y}_i, \cdot)$ is surjective, and from this we have

$$\max_{\mathbf{h}_i} \Pr(\Xi = \xi(\mathbf{y}_i, \mathbf{h}_i) \mid \Psi = \psi(\mathbf{y}_i), \tau, t) = \Pr(\Xi = \xi_j \mid \Psi = \tilde{y}_j, \tau, t).$$

Site split likelihood

Let ξ_j be such a choice for each $1 \leq j \leq q$. Then, the likelihood in (3) written as a product over site patterns as opposed to sites is

$$\begin{aligned}
L'_n(\tau, t; \mathbf{Y}) &= \max_{\mathbf{H}} \; L_n(\tau, t; \mathbf{Y}, \mathbf{H}) \\
&= \prod_{i=1}^{n} \max_{\mathbf{h}_i} \; \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) \\
&\propto \prod_{i=1}^{n} \max_{\mathbf{h}_i} \; \Pr(\Psi = \psi(\mathbf{y}_i) \mid \tau, t) \cdot \Pr(\Xi = \xi(\mathbf{y}_i, \mathbf{h}_i) \mid \Psi = \psi(\mathbf{y}_i), \tau, t) \\
&= \prod_{i=1}^{n} \Pr(\Psi = \psi(\mathbf{y}_i) \mid \tau, t) \cdot \max_{\mathbf{h}_i} \Pr(\Xi = \xi(\mathbf{y}_i, \mathbf{h}_i) \mid \Psi = \psi(\mathbf{y}_i), \tau, t) \\
&= \prod_{j=1}^{q} \left[\Pr(\Psi = \tilde{y}_j \mid \tau, t) \cdot \Pr(\Xi = \xi_j \mid \Psi = \tilde{y}_j, \tau, t)\right]^{n_j(\mathbf{Y})} \qquad (5) \quad \texttt{\{eq:site_pattern_lik}
\end{aligned}$$

where $n_j(\mathbf{Y})$ is the number of columns in $\mathbf{Y}$ that project to site split $\tilde{y}_j$.

Let

$$L''_n(\tau, t; \mathbf{Y}) = \prod_{j=1}^{q} \left[\Pr(\Psi = \tilde{y}_j \mid \tau, t) \cdot \Pr(\Xi = \xi_j \mid \Psi = \tilde{y}_j, \tau, t)\right]^{n_j(\mathbf{Y})}$$

be the final product in (5). Assume n observations are generated from a

model with parameters (τ^*, t^*). We have

$$\begin{aligned}
&\frac{1}{n} \log L''_n(\tau, t; \mathbf{Y}) \\
&\quad = \sum_{j=1}^{q} \frac{n_j(\mathbf{Y})}{n} \cdot \log \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, t) \\
&\quad = \sum_{j=1}^{q} \frac{n_j(\mathbf{Y})}{n} \cdot [\log \Pr(\Psi = \tilde{y}_j \mid \tau, t) + \log \Pr(\Xi = \xi_j \mid \Psi = \tilde{y}_j, \tau, t)]
\end{aligned}$$

so that, in the $n \to \infty$ limit,

$$\begin{aligned}
&\frac{1}{n} \log L''_n(\tau, t; \mathbf{Y}) \\
&\quad \to \sum_{j=1}^{q} \Pr(\Psi = \tilde{y}_j \mid \tau^*, t^*) \cdot [\log \Pr(\Psi = \tilde{y}_j \mid \tau, t) + \log \Pr(\Xi = \xi_j \mid \Psi = \tilde{y}_j, \tau, t)].
\end{aligned} \tag{6}$$

{eq:site_pattern_pro

Define the divergence quantity

$$D_{\tau^*, t^*}(\tau, t) = \sum_{j=1}^{q} \Pr(\Psi = \tilde{y}_j \mid \tau^*, t^*) \cdot \log \Pr(\Psi = \tilde{y}_j \mid \tau, t)$$

and the partial log-likelihood

$$\tilde{\ell}_{\tau^*, t^*}(\tau, t) = \sum_{j=1}^{q} \Pr(\Psi = \tilde{y}_j \mid \tau^*, t^*) \cdot \log \Pr(\Xi = \xi_j \mid \Psi = \tilde{y}_j, \tau, t)$$

so that (6) is

$$\ell_{\tau^*, t^*}(\tau, t) = D_{\tau^*, t^*}(\tau, t) + \tilde{\ell}_{\tau^*, t^*}(\tau, t). \tag{7}$$

{eq:log_likelihood_s

Hadamard representation

We state the Hadamard representation of site split generating probabilities, following Section 8.6 of Semple and Steel [2003]. For each edge e define the

edge "fidelity" for that edge as

$$\theta(e) = 1 - 2p(e).$$

For an even-sized subset of $Y \subseteq \mathcal{S}$, we define the path set $P(Y)$ as the set of edges in the path connecting both elements of Y. For n taxa, the probability of observing site split $A \in \mathcal{P}(\mathcal{Y})$ is

$$p_A = \frac{1}{2^{n-1}} \sum_{Y \subseteq \mathcal{S}: |Y| \equiv 0 (\mathrm{mod}\ 2)} \left[(-1)^{|Y \cap A|} \prod_{e \in P(Y)} \theta(e) \right]. \tag{8}$$ {eq:hadamard_probabi

By convention, we set $P(\emptyset) = \emptyset$ and $\prod_{e \in \emptyset} \theta(e) = 1$. For notational convenience, let

$$p_{\tilde{y}_j} := \Pr(\Psi = \tilde{y}_j \mid \tau_1, t),$$

for any site split $\tilde{y}_j$. Table S1 contains calculations of site pattern probabilities for our two topologies (Fig. 1).

$\tilde{y}_j$	$p_{\tilde{y}_j}$	$\Pr(\Psi = \tilde{y}_j \mid \tau_1, t)$	$\Pr(\Psi = \tilde{y}_j \mid \tau_2, t)$
$\emptyset$	$p_\emptyset$	$1 + x^2 + y^2 + 4xyw + x^2y^2$	$1 + 2xy + 2xyw + x^2w + y^2w + x^2y^2$
$\{1\}$	p_1	$1 - x^2 + y^2 - x^2y^2$	$1 - x^2w + y^2w - x^2y^2$
$\{2\}$	p_2	$1 + x^2 - y^2 - x^2y^2$	$1 + x^2w - y^2w - x^2y^2$
$\{3\}$	p_3	$1 - x^2 + y^2 - x^2y^2$	$1 - x^2w + y^2w - x^2y^2$
$\{1,2\}$	p_{12}	$1 - x^2 - y^2 + x^2y^2$	$1 + 2xy - 2xyw - x^2w - y^2w + x^2y^2$
$\{1,3\}$	p_{13}	$1 + x^2 + y^2 - 4xyw + x^2y^2$	$1 - 2xy - 2xyw + x^2w + y^2w + x^2y^2$
$\{2,3\}$	p_{23}	$1 - x^2 - y^2 + x^2y^2$	$1 - 2xy + 2xyw - x^2w - y^2w + x^2y^2$
$\{1,2,3\}$	p_{123}	$1 + x^2 - y^2 - x^2y^2$	$1 + x^2w - y^2w - x^2y^2$

Table S1: Site pattern probabilities $p_{\tilde{y}_j}$ on the Farris tree τ_1 and the Felsenstein tree τ_2 obtained using the Hadamard transform. All values multiplied by $1/8$. {tab:sitepatprob}

Example

We follow with an expository example computing these probabilities and likelihoods. Consider the fixed, binary four-taxon tree τ_1 in Fig. 1a—this is commonly known as the "Farris zone" topology. The set of all possible character assignments is

$$\begin{aligned}\mathcal{P}(\{1,2,3,4\}) = \{&\emptyset, \{1,2,3,4\}, \{1\}, \{2,3,4\}, \{2\}, \{1,3,4\}, \{3\}, \{1,2,4\},\\ &\{1,2\}, \{3,4\}, \{1,3\}, \{2,4\}, \{2,3\}, \{1,4\}, \{1,2,3\}, \{1,4\}\}.\end{aligned}$$

where each set indicates the tips assigned the character 1. For example, $\emptyset$ is the labeling 0000 and $\{1,3,4\}$ is the labeling 1011. Symmetry allows us to group adjacent pairs in $\mathcal{P}(\{1,2,3,4\})$ into equiprobable splits, letting $\mathcal{Y} = \{1,2,3\}$. The unique site splits, collapsing complements, are

$$\begin{aligned}\mathcal{P}(\mathcal{Y}) &= \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}\\ &:= \{\tilde{y}_1, \ldots, \tilde{y}_8\}.\end{aligned}$$

Since we identify character complements, we do not consider the additional splits

$$\begin{aligned}&\mathcal{P}(\{1,2,3,4\}) \setminus \mathcal{P}(\mathcal{Y}) =\\ &\quad \{\{1,2,3,4\}, \{2,3,4\}, \{1,3,4\}, \{1,2,4\}, \{3,4\}, \{2,4\}, \{1,4\}, \{4\}\},\end{aligned}$$

the symmetry of the binary character model allowing us to focus only on the elements of $\mathcal{P}(\mathcal{Y})$. This tree has two internal nodes with $\mathcal{H} = \{1,2\}$ and unique ancestral state splits

$$\mathcal{P}(\mathcal{H}) = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}.$$

Internal node $\{1\}$ is the node connected to leaves $\{1\}$ and $\{3\}$ and internal node $\{2\}$ connected to leaves $\{2\}$ and $\{4\}$. The mapping from characters to splits in this case will depend on the characters at the tips and the ancestral states. For example, we take both $\psi(0000) = \emptyset$ and $\psi(1111) = \emptyset$.

Similarly, we have $\xi(0000, 00) = \emptyset$ and $\xi(1111, 11) = \emptyset$, needing to take the complement of all the characters present on the tree to identify splits. We cannot identify complements for ancestral states in the same way as tip states since, for $\tilde{y} \in \mathcal{P}(\mathcal{Y})$,

$$\Pr(\Xi = \emptyset \mid \Psi = \tilde{y}, \tau, t) \neq \Pr(\Xi = \{1, 2\} \mid \Psi = \tilde{y}, \tau, t)$$

in general.

For each site split $\tilde{y} \in \mathcal{P}(\mathcal{Y})$, we maximize the likelihood over all $\tilde{h} \in \mathcal{P}(\mathcal{H})$. A maximum occurs at one of possibly several ancestral splits in $\mathcal{P}(\mathcal{H})$, defined via $\eta_j(\tau, t)$ for the jth site split. As a simple example, say all branch lengths correspond to a probability p ($< 1/2$) of changing character along that branch, with $t = \{p, p, p, p, p\}$. The probabilities of observing ancestral splits for $\tilde{y}_1 = \emptyset$ are

$$\Pr(\Xi = \emptyset \mid \Psi = \emptyset, \tau, t) = (1 - p)^5,$$

$$\Pr(\Xi = \{1\} \mid \Psi = \emptyset, \tau, t) = \Pr(\Xi = \{2\} \mid \Psi = \emptyset, \tau, t) = p^3(1 - p)^2,$$

$$\Pr(\Xi = \{1, 2\} \mid \Psi = \emptyset, \tau, t) = p^4(1 - p).$$

The set of most likely ancestral states contains a single element, here $\eta_1(\tau, t) = \{\emptyset\}$. Then, taking $\xi_1 \in \eta_1(\tau, t)$ we have

$$\Pr(\Xi = \xi_1 \mid \Psi = \emptyset, \tau, t) = \Pr(\Xi = \emptyset \mid \Psi = \emptyset, \tau, t) = (1 - p)^5.$$

For $\tilde{y}_5 = \{1, 2\}$ we have

$$\Pr(\Xi = \emptyset \mid \Psi = \{1, 2\}, \tau, t) = \Pr(\Xi = \{1, 2\} \mid \Psi = \{1, 2\}, \tau, t) = p^2(1 - p)^3,$$

$$\Pr(\Xi = \{1\} \mid \Psi = \{1, 2\}, \tau, t) = \Pr(\Xi = \{2\} \mid \Psi = \{1, 2\}, \tau, t) = p^3(1 - p)^2.$$

Here, the set of most likely ancestral states is $\eta_5(\tau, t) = \{\emptyset, \{1, 2\}\}$, and, for $\xi_5 \in \eta_5(\tau, t)$,

$$\Pr(\Xi = \xi_5 \mid \Psi = \{1, 2\}, \tau, t) = p^2(1 - p)^3.$$

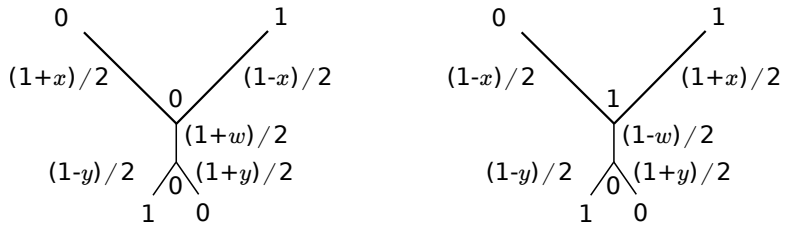


Figure S1: Example likelihood computations on the Farris tree τ_1 for fidelities x, y, and w. Edges labeled by the probability of substitution along that edge. In (a), we compute the product to obtain $\Pr(\mathbf{h} = \emptyset \mid \mathbf{y} = \{2,3\}, \tau_1, t) = (1+x)(1-x)(1+y)(1-y)(1+w)/32$. In (b), the same process yields $\Pr(\mathbf{h} = \{1\} \mid \mathbf{y} = \{2,3\}, \tau_1, t) = (1+x)(1-x)(1+y)(1-y)(1-w)/32$. {fig:example_likelih

Likelihood computations

To compute the likelihood of observing a set of data, we need $\Pr(\mathbf{h} = \tilde{h}_k \mid \mathbf{y} = \tilde{y}_j, \tau, t)$ for each $\tilde{h}_k$ and $\tilde{y}_j$. Using branch fidelities, the probability of a character change along a branch with fidelity parameter θ is $(1-\theta)/2$, while the probability of a character remaining the same is $(1+\theta)/2$. See Fig. S1 for the parameters on an example site pattern on the Farris tree. Likelihood computations for all site patterns and ancestral states are in Tables S2 and S3. Taking maxima row-wise of each table results in Table S4.

Form of the likelihood

Consider the Farris tree with arbitrary fidelities, i.e., $\tilde{t} = \{x_1, y_1, x_2, y_2, w\}$. We now show that, in the case of the Farris tree, exchanging x_1 with x_2 and y_1 with y_2 does not change the value of the likelihood, and that constraining both the top two branch parameters to be equal and the bottom two branch parameters to be equal during inference obtains the same maximum likelihood estimate as in the case of arbitrary branch parameters. Using the Hadamard transform, we calculate the generating probabilities

$\Pr(\mathbf{h} = \tilde{h}_k \mid \mathbf{y} = \tilde{y}_j, \tau_1, t)$

	$\tilde{h}_k$	
$\tilde{y}_j$	$\emptyset$	$\{2\}$
$\emptyset$	$(1+x)^2(1+w)(1+y)^2$	$(1+x)^2(1-w)(1-y)^2$
$\{1\}$	$(1+x)(1-x)(1+w)(1+y)^2$	$(1+x)(1-x)(1-w)(1-y)^2$
$\{2\}$	$(1+x)^2(1+w)(1+y)(1-y)$	$(1+x)^2(1-w)(1+y)(1-y)$
$\{3\}$	$(1+x)(1-x)(1+w)(1+y)^2$	$(1+x)(1-x)(1-w)(1-y)^2$
$\{1,2\}$	$(1+x)(1-x)(1+w)(1+y)(1-y)$	$(1+x)(1-x)(1-w)(1+y)(1-y)$
$\{1,3\}$	$(1-x)^2(1+w)(1+y)^2$	$(1-x)^2(1-w)(1-y)^2$
$\{2,3\}$	$(1+x)(1-x)(1+w)(1+y)(1-y)$	$(1+x)(1-x)(1-w)(1+y)(1-y)$
$\{1,2,3\}$	$(1-x)^2(1+w)(1+y)(1-y)$	$(1-x)^2(1-w)(1+y)(1-y)$
	$\{1\}$	$\{1,2\}$
$\emptyset$	$(1-x)^2(1-w)(1+y)^2$	$(1-x)^2(1+w)(1-y)^2$
$\{1\}$	$(1+x)(1-x)(1-w)(1+y)^2$	$(1+x)(1-x)(1+w)(1-y)^2$
$\{2\}$	$(1-x)^2(1-w)(1+y)(1-y)$	$(1-x)^2(1+w)(1+y)(1-y)$
$\{3\}$	$(1+x)(1-x)(1-w)(1+y)^2$	$(1+x)(1-x)(1+w)(1-y)^2$
$\{1,2\}$	$(1+x)(1-x)(1-w)(1+y)(1-y)$	$(1+x)(1-x)(1+w)(1+y)(1-y)$
$\{1,3\}$	$(1+x)^2(1-w)(1+y)^2$	$(1+x)^2(1+w)(1-y)^2$
$\{2,3\}$	$(1+x)(1-x)(1-w)(1+y)(1-y)$	$(1+x)(1-x)(1+w)(1+y)(1-y)$
$\{1,2,3\}$	$(1+x)^2(1-w)(1+y)(1-y)$	$(1+x)^2(1+w)(1+y)(1-y)$

Table S2: Likelihood calculations for all site patterns $\tilde{y}_j$ and internal states $\tilde{h}_k$ of the Farris tree τ_1. All values multiplied by $1/32$.

{tab:farris_likelihc

$\Pr(\mathbf{h} = \tilde{h}_k \mid \mathbf{y} = \tilde{y}_j, \tau_2, t)$

	$\tilde{h}_k$	
$\tilde{y}_j$	$\emptyset$	$\{2\}$
$\emptyset$	$(1+x)^2(1+w)(1+y)^2$	$(1+x)(1-x)(1-w)(1+y)(1-y)$
$\{1\}$	$(1+x)(1-x)(1+w)(1+y)^2$	$(1-x)^2(1-w)(1+y)(1-y)$
$\{2\}$	$(1+x)^2(1+w)(1+y)(1-y)$	$(1+x)(1-x)(1-w)(1-y)^2$
$\{3\}$	$(1+x)(1-x)(1+w)(1+y)^2$	$(1+x)^2(1-w)(1+y)(1-y)$
$\{1,2\}$	$(1+x)(1-x)(1+w)(1+y)(1-y)$	$(1-x)^2(1-w)(1-y)^2$
$\{1,3\}$	$(1-x)^2(1+w)(1+y)^2$	$(1+x)(1-x)(1-w)(1+y)(1-y)$
$\{2,3\}$	$(1+x)(1-x)(1+w)(1+y)(1-y)$	$(1+x)^2(1-w)(1-y)^2$
$\{1,2,3\}$	$(1-x)^2(1+w)(1+y)(1-y)$	$(1+x)(1-x)(1-w)(1-y)^2$
	$\{1\}$	$\{1,2\}$
$\emptyset$	$(1+x)(1-x)(1-w)(1+y)(1-y)$	$(1-x)^2(1+w)(1-y)^2$
$\{1\}$	$(1+x)^2(1-w)(1+y)(1-y)$	$(1+x)(1-x)(1+w)(1-y)^2$
$\{2\}$	$(1+x)(1-x)(1-w)(1+y)^2$	$(1-x)^2(1+w)(1+y)(1-y)$
$\{3\}$	$(1-x)^2(1-w)(1+y)(1-y)$	$(1+x)(1-x)(1+w)(1-y)^2$
$\{1,2\}$	$(1+x)^2(1-w)(1+y)^2$	$(1+x)(1-x)(1+w)(1+y)(1-y)$
$\{1,3\}$	$(1+x)(1-x)(1-w)(1+y)(1-y)$	$(1+x)^2(1+w)(1-y)^2$
$\{2,3\}$	$(1-x)^2(1-w)(1+y)^2$	$(1+x)(1-x)(1+w)(1+y)(1-y)$
$\{1,2,3\}$	$(1+x)(1-x)(1-w)(1+y)^2$	$(1+x)^2(1+w)(1+y)(1-y)$

Table S3: Likelihood calculations for all site patterns $\tilde{y}_j$ and internal states $\tilde{h}_k$ of the Felsenstein tree τ_2. All values multiplied by $1/32$.

{tab:fels_likelihood

Farris tree ($\tau = \tau_1$)

$\tilde{y}_j$	$\eta_j(\tau, t)$	$\Pr(\Xi = \xi_j \mid \Psi = \tilde{y}_j, \tau, t)$
$\emptyset$	$\emptyset$	$(1+x)^2(1+w)(1+y)^2$
$\{1\}$	$\emptyset$	$(1+x)(1-x)(1+w)(1+y)^2$
$\{2\}$	$\emptyset$	$(1+x)^2(1+w)(1+y)(1-y)$
$\{3\}$	$\emptyset$	$(1+x)(1-x)(1+w)(1+y)^2$
$\{1,2\}$	$\{\emptyset, \{1,2\}\}$	$(1+x)(1-x)(1+w)(1+y)(1-y)$
$\{1,3\}$	$\emptyset$	$(1-x)^2(1+w)(1+y)^2$
	$\{1\}$	$(1+x)^2(1-w)(1+y)^2$
	$\{1,2\}$	$(1+x)^2(1+w)(1-y)^2$
$\{2,3\}$	$\{\emptyset, \{1,2\}\}$	$(1+x)(1-x)(1+w)(1+y)(1-y)$
$\{1,2,3\}$	$\{1,2\}$	$(1+x)^2(1+w)(1+y)(1-y)$

Felsenstein tree ($\tau = \tau_2$)

$\tilde{y}_j$	$\eta_j(\tau, t)$	$\Pr(\Xi = \xi_j \mid \Psi = \tilde{y}_j, \tau, t)$
$\emptyset$	$\emptyset$	$(1+x)^2(1+w)(1+y)^2$
$\{1\}$	$\emptyset$	$(1+x)(1-x)(1+w)(1+y)^2$
	$\{1\}$	$(1+x)^2(1-w)(1+y)(1-y)$
$\{2\}$	$\emptyset$	$(1+x)^2(1+w)(1+y)(1-y)$
	$\{1\}$	$(1+x)(1-x)(1-w)(1+y)^2$
$\{3\}$	$\emptyset$	$(1+x)(1-x)(1+w)(1+y)^2$
	$\{2\}$	$(1+x)^2(1-w)(1+y)(1-y)$
$\{1,2\}$	$\{\emptyset, \{1,2\}\}$	$(1+x)(1-x)(1+w)(1+y)(1-y)$
	$\{1\}$	$(1+x)^2(1-w)(1+y)^2$
$\{1,3\}$	$\emptyset$	$(1-x)^2(1+w)(1+y)^2$
	$\{\{1\}, \{2\}\}$	$(1+x)(1-x)(1-w)(1+y)(1-y)$
	$\{1,2\}$	$(1+x)^2(1+w)(1-y)^2$
$\{2,3\}$	$\{\emptyset, \{1,2\}\}$	$(1+x)(1-x)(1+w)(1+y)(1-y)$
	$\{1\}$	$(1-x)^2(1-w)(1+y)^2$
	$\{2\}$	$(1+x)^2(1-w)(1-y)^2$
$\{1,2,3\}$	$\{2\}$	$(1+x)(1-x)(1-w)(1+y)^2$
	$\{1,2\}$	$(1+x)^2(1+w)(1+y)(1-y)$

Table S4: Likelihood calculations for all site patterns $\tilde{y}_j$ and ancestral state partitions η_j after maximizing over ancestral states on the Farris tree τ_1 and the Felsenstein tree τ_2. All values multiplied by $1/32$. Likelihoods with multiple entries have maxima determined by unknown branch length parameters. See Tables S2 and S3 for full calculations.

{tab:likelihoods}

on the Farris tree. For site split $\emptyset$,

$$\begin{aligned}\Pr(\Psi = \emptyset \mid \tau_1, \tilde{t}) &= \frac{1}{8}(1 + x_1x_2 + y_1y_2 + x_1y_1w + x_1y_2w + y_1x_2w + x_2y_2w + x_1y_1x_2y_2) \\ &= \frac{1}{8}(1 + x_1x_2 + y_1y_2 + w[x_1y_1 + x_1y_2 + y_1x_2 + x_2y_2] + x_1y_1x_2y_2) \\ &= \frac{1}{8}(1 + x_1x_2 + y_1y_2 + w[x_1 + x_2][y_1 + y_2] + x_1y_1x_2y_2).\end{aligned}$$

and this probability is unchanged when x_1 is exchanged with x_2 and y_1 is exchanged with y_2. All other generating probabilities will differ only in the signs of each term. For example, for site split $\{1\}$ we have

$$\Pr(\Psi = \{1\} \mid \tau_1, \tilde{t}) = \frac{1}{8}(1 - x_1x_2 + y_1y_2 + w[-x_1 + x_2][y_1 + y_2] - x_1y_1x_2y_2)$$

and for site split $\{3\}$ we have

$$\Pr(\Psi = \{3\} \mid \tau_1, \tilde{t}) = \frac{1}{8}(1 - x_1x_2 + y_1y_2 + w[x_1 - x_2][y_1 + y_2] - x_1y_1x_2y_2)$$

meaning if we exchange the values of x_1 and x_2 then these probabilities swap values. The corresponding possibilities for the likelihood values are

$$\begin{aligned}\Pr(\Xi = \emptyset \mid \Psi = \{1\}, \tau_1, \tilde{t}) &= \frac{1}{32}(1 - x_1)(1 + x_2)(1 + w)(1 + y_1)(1 + y_2); \\ \Pr(\Xi = \{1\} \mid \Psi = \{1\}, \tau_1, \tilde{t}) &= \frac{1}{32}(1 + x_1)(1 - x_2)(1 - w)(1 + y_1)(1 + y_2); \\ \Pr(\Xi = \{2\} \mid \Psi = \{1\}, \tau_1, \tilde{t}) &= \frac{1}{32}(1 - x_1)(1 + x_2)(1 - w)(1 - y_1)(1 - y_2); \\ \Pr(\Xi = \{1,2\} \mid \Psi = \{1\}, \tau_1, \tilde{t}) &= \frac{1}{32}(1 + x_1)(1 - x_2)(1 + w)(1 - y_1)(1 - y_2);\end{aligned}$$

for site split $\{1\}$ and

$$\begin{aligned}\Pr(\Xi = \emptyset \mid \Psi = \{3\}, \tau_1, \tilde{t}) &= \frac{1}{32}(1 + x_1)(1 - x_2)(1 + w)(1 + y_1)(1 + y_2); \\ \Pr(\Xi = \{1\} \mid \Psi = \{3\}, \tau_1, \tilde{t}) &= \frac{1}{32}(1 - x_1)(1 + x_2)(1 - w)(1 + y_1)(1 + y_2); \\ \Pr(\Xi = \{2\} \mid \Psi = \{3\}, \tau_1, \tilde{t}) &= \frac{1}{32}(1 + x_1)(1 - x_2)(1 - w)(1 - y_1)(1 - y_2);\end{aligned}$$

$$\Pr(\Xi = \{1,2\} \mid \Psi = \{3\}, \tau_1, \tilde{t}) = \frac{1}{32}(1-x_1)(1+x_2)(1+w)(1-y_1)(1-y_2);$$

for site split $\{3\}$, which also both swap values when x_1 and x_2 are exchanged.

The same can be done for the splits $\{2\}$ and $\{1,2,3\}$ by exchanging y_1 and y_2 as well as $\{1,2\}$ and $\{1,3\}$ by exchanging both x_1 with x_2 and y_1 with y_2. The split $\{1,3\}$ is unchanged by exchanging x_1 with x_2 and y_1 with y_2.

Since exchanging x_1 and x_2 does not change the value of the log-likelihood $\ell_{\tau_1,t^*}(\tau_1,\tilde{t})$, if there is a unique maximum of the log-likelihood we will have $x_1 = x_2$ at the maximum. An analogous statement holds for y_1 and y_2. It remains to show that the joint inference procedure maximizing (3) results in a unique estimate of t. Without loss of generality, we focus on the single parameter x_1 and show we obtain a unique maximum when performing joint inference; similar arguments hold for the remaining parameters. We decompose the likelihood into the entropy and partial likelihood terms as in (7). By Gibbs's inequality, the entropy term $D_{\tau^*,t^*}(\tau^*,t)$ has a unique maximum over t—namely t^*. For the general case of $\tilde{t} = \{x_1, y_2, x_1, y_2, w\}$, we can compute the partial likelihood $\tilde{\ell}_{\tau^*,t^*}(\tau^*,t)$ as in Table S2. Clearly this partial likelihood will be of the form $g(x_1) = p\log(1+x_1)+(1-p)\log(1-x_1)$ with $0 \leq p \leq 1$, which, by the second derivative test, is concave for $x_1 \in [0,1]$ and thus has a unique maximum. Since both summands of the likelihood are concave on $x_1 \in [0,1]$, the likelihood is as well, and there exists a single $x_1 \in [0,1]$ maximizing the likelihood. Given this and the symmetry of the likelihood, we let $x_1 = x_2 = x$ and $y_1 = y_2 = y$. The Felsenstein tree does not admit this property, but, since we are interested in a lower bound for this tree, we simplify the objective function by constraining $x_1 = x_2$ and $y_1 = y_2$ similarly.

Theorems and proofs

Theorem 1. *Let* $t^* = \{x^*, y^*, x^*, y^*, y^*\}$ *and* $t = \{x, y, x, y, w\}$. *There exist* $C_0(x^*, y^*), C_1(x^*, y^*)$, *and a set of* $0 < x^*, y^* < 1$ *such that*

$$\max_t \ \ell_{\tau_1, t^*}(\tau_1, t) \leq C_0(x^*, y^*),$$

$$C_1(x^*, y^*) \leq \max_t \ \ell_{\tau_1, t^*}(\tau_2, t)$$

with $C_0(x^*, y^*) < C_1(x^*, y^*)$.

Proof. In general,

$$\max_t \ \ell_{\tau^*, t^*}(\tau, t) \leq D_{\tau^*, t^*}(\tau^*, t^*) + \tilde{\ell}_{\tau^*, t^*}(\tau, \hat{t})$$

is an upper bound for the joint maximum of (7) using Gibbs's inequality

$$D_{\tau^*, t^*}(\tau, t) \leq D_{\tau^*, t^*}(\tau^*, t^*)$$

and

$$\hat{t} = \operatorname*{argmax}_t \ \tilde{\ell}_{\tau^*, t^*}(\tau, t).$$

Similarly,

$$\max_t \ \ell_{\tau^*, t^*}(\tau, t) \geq D_{\tau^*, t^*}(\tau, \hat{t}) + \tilde{\ell}_{\tau^*, t^*}(\tau, \hat{t})$$

is a lower bound.

Assume $\tau^* = \tau_1$. The Farris tree log-likelihood takes one of three values depending on branch lengths, which is due to site split $\{1, 3\}$ (see the upper table of Table S4).

We write out the case for the ancestral state split $\{1\}$, and show the other two cases follow a similar argument. Directly substituting calculations from Table S1 and Table S4 where $\tau = \tau_1$ in (6), the log-likelihood is, suppressing normalizing constants $-\log 8$ (from the first additive term of (6)) and $-\log 32$ (from the second),

$$\begin{aligned} \ell_{\tau_1, t^*}(\tau_1, t) = p_\emptyset \cdot \log(1 + x^2 + y^2 + 4xyw + x^2y^2) \\ + p_1 \cdot \log(1 - x^2 + y^2 - x^2y^2) \end{aligned}$$

$$
\begin{aligned}
&+ p_2 \cdot \log(1 + x^2 - y^2 - x^2y^2) \\
&+ p_3 \cdot \log(1 - x^2 + y^2 - x^2y^2) \\
&+ p_{12} \cdot \log(1 - x^2 - y^2 + x^2y^2) \\
&+ p_{13} \cdot \log(1 + x^2 + y^2 - 4xyw + x^2y^2) \\
&+ p_{23} \cdot \log(1 - x^2 - y^2 + x^2y^2) \\
&+ p_{123} \cdot \log(1 + x^2 - y^2 - x^2y^2) \\
&+ p_{\emptyset} \cdot \log((1 + x)^2(1 + w)(1 + y)^2) \\
&+ p_1 \cdot \log((1 + x)(1 - x)(1 + w)(1 + y)^2) \\
&+ p_2 \cdot \log((1 + x)^2(1 + w)(1 + y)(1 - y)) \\
&+ p_3 \cdot \log((1 + x)(1 - x)(1 + w)(1 + y)^2) \\
&+ p_{12} \cdot \log((1 + x)(1 - x)(1 + w)(1 + y)(1 - y)) \\
&+ p_{13} \cdot \log((1 + x)^2(1 - w)(1 + y)^2) \\
&+ p_{23} \cdot \log((1 + x)(1 - x)(1 + w)(1 + y)(1 - y)) \\
&+ p_{123} \cdot \log((1 + x)^2(1 + w)(1 + y)(1 - y)). \qquad (9) \quad \{\texttt{eq:farris\_likelihoc}
\end{aligned}
$$

Our plan is to simplify the likelihood to remove log-of-quadratic terms of x, y, and w, obtaining a likelihood that has a closed-form maximum in each variable. To bound the generating probabilities, we use the facts that, for $x, y \in [0, 1]$,

$$
\begin{aligned}
p_{12} = p_{23} &= 1 - x^2 - y^2 + x^2y^2 = (1 + x)(1 - x)(1 + y)(1 - y) \\
p_1 = p_3 &= 1 - x^2 + y^2 - x^2y^2 = (1 + x)(1 - x)(1 + y^2) \leq (1 + x)(1 - x)(1 + y) \\
p_2 = p_{123} &= 1 + x^2 - y^2 - x^2y^2 = (1 + x^2)(1 + y)(1 - y) \leq (1 + x)(1 + y)(1 - y).
\end{aligned}
$$

To bound the remaining $p_{\emptyset}$ and p_{13}, we use

$$
\begin{aligned}
1 + x^2 + y^2 + x^2y^2 &= (1 + x^2)(1 + y^2) \leq (1 + x)(1 + y) \\
4xy = 2x \cdot 2y &\leq (1 + x^2)(1 + y^2) \leq (1 + x)(1 + y)
\end{aligned}
$$

so that the $p_\emptyset$ term is bounded as

$$\begin{aligned} p_\emptyset &= \log(1 + x^2 + y^2 + 4xyw + x^2y^2) \\ &\leq \log(1 + x^2 + y^2 + 4xy + x^2y^2) \\ &\leq \log(2(1 + x)(1 + y)) \end{aligned}$$

and p_{13} is bounded as

$$\begin{aligned} p_{13} &= \log(1 + x^2 + y^2 - 4xyw + x^2y^2) \\ &\leq \log(1 + x^2 + y^2 + x^2y^2) \\ &\leq \log((1 + x)(1 + y)). \end{aligned}$$

Factoring and making these substitutions (again, under the assumption that the ancestral state split is $\{1\}$) results in

$$\begin{aligned} &\ell_{\tau_1,t^*}(\tau_1, t) \\ &\quad\leq p_\emptyset \cdot \log(2(1 + x)(1 + y)) + p_1 \cdot \log((1 + x)(1 - x)(1 + y)) \\ &\qquad + p_2 \cdot \log((1 + x)(1 + y)(1 - y)) + p_3 \cdot \log((1 + x)(1 - x)(1 + y)) \\ &\qquad + p_{12} \cdot \log((1 + x)(1 - x)(1 + y)(1 - y)) + p_{13} \cdot \log((1 + x)(1 + y)) \\ &\qquad + p_{23} \cdot \log((1 + x)(1 - x)(1 + y)(1 - y)) + p_{123} \cdot \log((1 + x)(1 + y)(1 - y)) \\ &\qquad + p_\emptyset \cdot \log((1 + x)^2(1 + w)(1 + y)^2) + p_1 \cdot \log((1 + x)(1 - x)(1 + w)(1 + y)^2) \\ &\qquad + p_2 \cdot \log((1 + x)^2(1 + w)(1 + y)(1 - y)) + p_3 \cdot \log((1 + x)(1 - x)(1 + w)(1 + y)^2) \\ &\qquad + p_{12} \cdot \log((1 + x)(1 - x)(1 + w)(1 + y)(1 - y)) + p_{13} \cdot \log((1 + x)^2(1 - w)(1 + y)^2) \\ &\qquad + p_{23} \cdot \log((1 + x)(1 - x)(1 + w)(1 + y)(1 - y)) + p_{123} \cdot \log((1 + x)^2(1 + w)(1 + y)(1 - y)) \\ &\quad= p_\emptyset \cdot \log(2) + \log(1 + x) + (p_1 + p_3 + p_{12} + p_{23}) \cdot \log(1 - x) \\ &\qquad + \log(1 + y) + (p_2 + p_{12} + p_{23} + p_{123}) \cdot \log(1 - y) \\ &\qquad + (2 - p_1 - p_3 - p_{12} - p_{23}) \cdot \log(1 + x) + (p_1 + p_3 + p_{12} + p_{23}) \cdot \log(1 - x) \\ &\qquad + (2 - p_2 - p_{12} - p_{23} - p_{123}) \cdot \log(1 + y) + (p_2 + p_{12} + p_{23} + p_{123}) \cdot \log(1 - y) \\ &\qquad + (1 - p_{13}) \cdot \log(1 + w) + p_{13} \cdot \log(1 - w) \\ &\quad= p_\emptyset \cdot \log(2) + (3 - p_1 - p_3 - p_{12} - p_{23}) \cdot \log(1 + x) + 2(p_1 + p_3 + p_{12} + p_{23}) \cdot \log(1 - x) \end{aligned}$$

$$+ (3 - p_2 - p_{12} - p_{23} - p_{123}) \cdot \log(1+y) + 2(p_2 + p_{12} + p_{23} + p_{123}) \cdot \log(1-y)$$
$$+ (1 - p_{13}) \cdot \log(1+w) + p_{13} \cdot \log(1-w).$$

412 To simplify, let

$$\begin{aligned} a_1 &= 3 - p_1 - p_3 - p_{12} - p_{23}, \\ a_2 &= 2(p_1 + p_3 + p_{12} + p_{23}), \\ a_3 &= 3 - p_2 - p_{12} - p_{23} - p_{123}, \\ a_4 &= 2(p_2 + p_{12} + p_{23} + p_{123}), \end{aligned} \tag{10}$$ {eq:a_const}

so that

$$\begin{aligned} &\ell_{\tau_1,t^*}(\tau_1, t) \\ &\quad \le p_\emptyset \cdot \log(2) + a_1 \cdot \log(1+x) + a_2 \cdot \log(1-x) + a_3 \cdot \log(1+y) \\ &\qquad + a_4 \cdot \log(1-y) + (1 - p_{13}) \cdot \log(1+w) + p_{13} \cdot \log(1-w). \end{aligned}$$

Maximizing over the unknown terms yields

$$\hat{x} = \frac{a_1 - a_2}{a_1 + a_2}, \ \hat{y} = \frac{a_3 - a_4}{a_3 + a_4}, \ \hat{w} = 1 - 2p_{13}.$$

Trivially $0 \le \hat{w} \le 1$ and $\hat{x}, \hat{y} \le 1$. We check that $\hat{x}, \hat{y} \ge 0$ to ensure these maxima are valid. For $\hat{x} \ge 0$ we need $a_1 \ge a_2$; letting

$$\tilde{p} = p_1 + p_3 + p_{12} + p_{23}$$

we have from (10) that $a_1 \ge a_2$ only if $3 \ge 3\tilde{p}$. Since $\tilde{p} = 1/8 \cdot (4 - 4(x^*)^2)$, we have $0 \le \tilde{p} \le 1/2$ implying $\hat{x} \ge 0$. The same approach works for $\hat{y}$. The

upper bound for the likelihood is then maximized at

$$\begin{aligned}
&\ell_{\tau_1,t^*}(\tau_1,t)\\
&\quad\leq p_\emptyset\cdot\log(2)+a_1\cdot\log\frac{2a_1}{a_1+a_2}+a_2\cdot\log\frac{2a_2}{a_1+a_2}+a_3\cdot\log\frac{2a_3}{a_3+a_4}\\
&\qquad+a_4\cdot\log\frac{2a_4}{a_3+a_4}+(1-p_{13})\cdot\log(2(1-p_{13}))+p_{13}\cdot\log(2p_{13})\\
&\quad:=C^1_{\tau_1,\tau_1}(x^*,y^*).
\end{aligned}\tag{11}\label{eq:farris-upper-bou}$$

The other two possible ancestral state splits for the likelihood of the site split $\{1,3\}$ admit similar simplifications. The upper bound in (11) is for the ancestral state split $\{1\}$. For the ancestral state split $\emptyset$ we see the upper bound will be the same except we lose a $(1+x)^2$ term, gain a $(1-x)^2$ term, lose a $(1-w)$ term, and gain a $(1+w)$ term. In this case, all terms involving w simplify to $\log(1+w)$, maximized at $\hat{w}=1$. For the constants above, a_1 is the multiplier for the $\log(1+x)$ term and a_2 for the $\log(1-x)$ term meaning that exchanging the above quadratic x terms yields new constants

$$\begin{aligned}
a'_1&=a_1-2p_{13},\\
a'_2&=a_2+2p_{13},
\end{aligned}\tag{12}\label{eq:a_const_prime_x}$$

for site split $\emptyset$ and

$$\begin{aligned}
a'_3&=a_3-2p_{13},\\
a'_4&=a_4+2p_{13},
\end{aligned}\tag{13}\label{eq:a_const_prime_y}$$

for site split $\{1,2\}$. To bound the case of $\eta_j(\tau_1,t)=\emptyset$,

$$\begin{aligned}
&\ell_{\tau_1,t^*}(\tau_1,t)\\
&\quad\leq p_\emptyset\cdot\log(2)+a'_1\cdot\log\frac{2a'_1}{a'_1+a'_2}+a'_2\cdot\log\frac{2a'_2}{a'_1+a'_2}\\
&\qquad+a_3\cdot\log\frac{2a_3}{a_3+a_4}+a_4\cdot\log\frac{2a_4}{a_3+a_4}+\log(2)\\
&\quad:=C^2_{\tau_1,\tau_1}(x^*,y^*),
\end{aligned}$$

and, for the case of $\eta_j(\tau_1, t) = \{1, 2\}$,

$$\begin{aligned}
&\ell_{\tau_1,t^*}(\tau_1, t) \\
&\quad \leq p_\emptyset \cdot \log(2) + a_1 \cdot \log \frac{2a_1}{a_1 + a_2} + a_2 \cdot \log \frac{2a_2}{a_1 + a_2} \\
&\qquad + a'_3 \cdot \log \frac{2a'_3}{a'_3 + a'_4} + a'_4 \cdot \log \frac{2a'_4}{a'_3 + a'_4} + \log(2) \\
&\quad := C^3_{\tau_1,\tau_1}(x^*, y^*).
\end{aligned}$$

We now need to check that

$$\hat{x} = \frac{a'_1 - a'_2}{a'_1 + a'_2}, \ \hat{y} = \frac{a'_3 - a'_4}{a'_3 + a'_4}$$

are both between zero and one. Using a similar argument as that where $\eta_j(\tau_1, t) = \{1, 3\}$, we only need to show $a'_1 \geq a'_2$, which is true if $3 - 2p_{13} \geq 3\tilde{p} + 2p_{13}$. Some rearranging and the fact that $0 \leq \tilde{p} \leq 1/2$ shows this is equivalent to showing $p_{13} \leq 3/8$. Using Table S1 and t^*,

$$p_{13} = \frac{1}{8}\left(1 + (x^*)^2 + (y^*)^2 - 4x^*(y^*)^2 + (x^*)^2(y^*)^2\right),$$

meaning we are interested in whether

$$1 + (x^*)^2 + (y^*)^2 - 4x^*(y^*)^2 + (x^*)^2(y^*)^2 \leq 3. \tag{14}$$ {eq:generating_ineq}

We see that

$$1+(x^*)^2+(y^*)^2-4x^*(y^*)^2+(x^*)^2(y^*)^2 \leq 3 \iff (y^*)^2\left(1 - 4x^* + (x^*)^2\right) \leq 2-(x^*)^2 \tag{15}$$ {eq:lenny}

and that

$$\begin{aligned}
(y^*)^2\left(1 - 4x^* + (x^*)^2\right) &\leq (y^*)^2\left(1 - 2x^* + (x^*)^2\right) \\
&= (y^*)^2(1 - x^*)^2 \\
&\leq (1 - x^*)^2 \\
&\leq 2 - (x^*)^2,
\end{aligned}$$

with the last inequality holding if $0 \leq x^* \leq 1$, showing that (15) holds. Thus, $p_{13} \leq 3/8$ and our $\hat{x}$ and $\hat{y}$ are valid.

All bounds are functions only of x, y through the true generating probabilities x^*, y^*. Call the upper bound

$$C_0(x^*, y^*) := \max\left(C^1_{\tau_1,\tau_1}(x^*, y^*), C^2_{\tau_1,\tau_1}(x^*, y^*), C^3_{\tau_1,\tau_1}(x^*, y^*)\right).$$

We construct a similar lower bound on the Felsenstein tree partial likelihood. In the Felsenstein case, there are many more optimal internal states depending on branch lengths than in the Farris case (see the lower table of Table S4). We first bound the entire likelihood below and then proceed with joint inference. To obtain a lower bound for the likelihood we replace $1 + w$ with $1 - w$ in Table S4 for $\tau = \tau_2$; in this case, we resolve many of the ambiguous likelihood terms. For example, for site split $\{1, 2\}$, after we replace the $1 + w$ term for $\eta_j(\tau, t) = \{\emptyset, \{1, 2\}\}$ with $1 - w$, we have

$$\max\left((1+x)^2(1-w)(1+y)^2, (1+x)(1-x)(1-w)(1+y)(1-y)\right)$$

where since

$$(1+x)^2(1-w)(1+y)^2 > (1+x)(1-x)(1-w)(1+y)(1-y),$$

the maximum for this site split can be bounded below by $(1+x)^2(1-w)(1+y)^2$. For the other site splits we consider two cases. If $(1+x)(1-y) > (1-x)(1+y)$, then

$$(1+x)^2(1-w)(1+y)(1-y) > (1+x)(1-x)(1-w)(1+y)^2$$

meaning the cases $\{1\}, \{2\}, \{3\}$ and $\{1, 2, 3\}$ will have likelihoods bounded below by

$$(1+x)^2(1-w)(1+y)(1-y).$$

For the cases $\{1, 3\}$ and $\{2, 3\}$ and the same condition, we have

$$(1+x)^2(1-w)(1-y)^2 > (1-x)^2(1-w)(1+y)^2$$

and

$$(1+x)^2(1-w)(1-y)^2 > (1+x)(1-x)(1-w)(1+y)(1-y),$$

yielding lower bounds of

$$(1+x)^2(1-w)(1-y)^2$$

in these cases. The partial likelihood in this case is then bounded below by

$$\begin{aligned}\tilde{\ell}_{\tau_1,t^*}(\tau_2,t) \geq\ & p_\emptyset \cdot \log((1+x)^2(1-w)(1+y)^2) \\ & + p_1 \cdot \log((1+x)^2(1-w)(1+y)(1-y)) \\ & + p_2 \cdot \log((1+x)^2(1-w)(1+y)(1-y)) \\ & + p_3 \cdot \log((1+x)^2(1-w)(1+y)(1-y)) \\ & + p_{12} \cdot \log((1+x)^2(1-w)(1+y)^2) \\ & + p_{123} \cdot \log((1+x)^2(1-w)(1+y)(1-y)) \\ & + p_{13} \cdot \log((1+x)^2(1-w)(1-y)^2) \\ & + p_{23} \cdot \log((1+x)^2(1-w)(1-y)^2).\end{aligned}$$

When $(1+x)(1-y) < (1-x)(1+y)$, similar arguments yield the lower bound

$$\begin{aligned}\tilde{\ell}_{\tau_1,t^*}(\tau_2,t) \geq\ & p_\emptyset \cdot \log((1+x)^2(1-w)(1+y)^2) \\ & + p_1 \cdot \log((1+x)(1-x)(1-w)(1+y)^2) \\ & + p_2 \cdot \log((1+x)(1-x)(1-w)(1+y)^2) \\ & + p_3 \cdot \log((1+x)(1-x)(1-w)(1+y)^2) \\ & + p_{12} \cdot \log((1+x)^2(1-w)(1+y)^2) \\ & + p_{123} \cdot \log((1+x)(1-x)(1-w)(1+y)^2) \\ & + p_{13} \cdot \log((1-x)^2(1-w)(1+y)^2) \\ & + p_{23} \cdot \log((1-x)^2(1-w)(1+y)^2).\end{aligned}$$

Letting

$$b = p_1 + p_2 + p_3 + p_{123}$$

yields

$$\tilde{\ell}_{\tau_1,t^*}(\tau_2, t) \geq 2 \cdot \log(1+x) + (2-b) \cdot \log(1+y) + b \cdot \log(1-y) + \log(1-w)$$

for the first lower bound, and we get the equivalent expression for the second lower bound by exchanging x and y. Maximizing either lower bound over t, we obtain

$$\tilde{\ell}_{\tau_1,t^*}(\tau_2, t) \geq 2 \cdot \log(2) + (2-b) \cdot \log(2-b) + b \cdot \log(b),$$

i.e., the maximum is achieved at $\{1, 1-b, 0\}$. Our lower bound is

$$C_1(x^*, y^*) := D_{\tau_1,t^*}(\tau_2, \{1, 1-b, 0\}) + 2 \cdot \log(2) + (2-b) \cdot \log(2-b) + b \cdot \log(b).$$

See Fig. 2 for the values of $0 < x^*, y^* < 1$ where $C_1(x^*, y^*) > C_0(x^*, y^*)$ and joint inference is inconsistent. □

Theorem 2. *Let*

$$\beta := \beta(x^*, y^*) = 1 + (x^*)^2 + (y^*)^2 + (x^*)^2(y^*)^2,$$

$$\gamma := \gamma(x^*, y^*) = 4x^*y^*.$$

The maximum likelihood value $\hat{w}$ is equal to 1 if

$$-\gamma^2 \left(1 + \frac{1}{2}\beta\right) + 2\gamma\beta x^*(y^*)^2 + \beta^2 \geq 0$$

and there exists a set of $0 < x^*, y^* < 1$ satisfying this.

Proof. Recalling Table S4 for the Farris tree $\tau = \tau_1$, the likelihood can take one of three forms corresponding to which ancestral state for $\tilde{y}_j = \{1, 3\}$ maximizes the likelihood over (x, y, w). If $x = x^*$ and $y = y^*$ are fixed, by collapsing like terms and ignoring terms with only x^* and y^*, the log-

likelihood in (9) as a function of w is

$$\begin{aligned}\ell_{\tau_1,t^*}(\tau_1, w) \sim\ & p_\emptyset \cdot \log(1 + (x^*)^2 + (y^*)^2 + 4x^*y^*w + (x^*)^2(y^*)^2) \\ & + p_{13} \cdot \log(1 + (x^*)^2 + (y^*)^2 - 4x^*y^*w + (x^*)^2(y^*)^2) \\ & + (1 - p_{13}) \cdot \log(1 + w) \\ & + p_{13} \cdot \log(1 - w)\end{aligned}$$

where $\sim$ denotes equality up to an additive constant. Similarly, the other two possibilities for the likelihood both satisfy

$$\begin{aligned}\ell_{\tau_1,t^*}(\tau_1, w) \sim\ & p_\emptyset \cdot \log(1 + (x^*)^2 + (y^*)^2 + 4x^*y^*w + (x^*)^2(y^*)^2) \\ & + p_{13} \cdot \log(1 + (x^*)^2 + (y^*)^2 - 4x^*y^*w + (x^*)^2(y^*)^2) \\ & + \log(1 + w).\end{aligned}$$

Since

$$\log(1 + w) \geq (1 - p_{13}) \cdot \log(1 + w) + p_{13} \cdot \log(1 - w)$$

with equality only when $w = 0$, the second likelihood is the likelihood with the maximum-likelihood ancestral state, and so we only analyze that case.

Now substitute in values for $p_\emptyset$ and p_{13} from Table S1. Simplifying, the likelihood can be written

$$\ell(w) = \alpha_1 \log(\beta + \gamma w) + \alpha_2 \log(\beta - \gamma w) + \log(1 + w)$$

where

$$\alpha_1 := \alpha_1(x^*, y^*) = \frac{1}{8}\left(1 + (x^*)^2 + (y^*)^2 + 4x^*(y^*)^2 + (x^*)^2(y^*)^2\right),$$

$$\alpha_2 := \alpha_2(x^*, y^*) = \frac{1}{8}\left(1 + (x^*)^2 + (y^*)^2 - 4x^*(y^*)^2 + (x^*)^2(y^*)^2\right),$$

$$\beta := \beta(x^*, y^*) = 1 + (x^*)^2 + (y^*)^2 + (x^*)^2(y^*)^2,$$

$$\gamma := \gamma(x^*, y^*) = 4x^*y^*.$$

Since x^* and y^* fall between zero and one, we make use of the inequalities

(excepting the cases of $x^* = y^* = 0$ and $x^* = y^* = 1$)

$$\alpha_1 > \alpha_2$$

and

$$\beta > \gamma.$$

The derivative of the log-likelihood with respect to w is

$$\ell'(w) := \frac{d}{dw}\ell(w) = \frac{\alpha_1\gamma}{\beta + \gamma w} - \frac{\alpha_2\gamma}{\beta - \gamma w} + \frac{1}{1+w}.$$

The inequality $\beta > \gamma$ implies that this function stays finite and that, when considering $\ell'(w) \lesseqgtr 0$, we equivalently consider $f(w) \lesseqgtr 0$ where f is the quadratic function

$$\begin{aligned} f(w) = (w)^2 \cdot (&-\gamma^2\alpha_1 - \gamma^2\alpha_2 - \gamma^2) \\ &+ w \cdot (\gamma\alpha_1\beta - \gamma^2\alpha_1 - \gamma\alpha_2\beta - \gamma^2\alpha_2) \\ &+ (\gamma\alpha_1\beta - \gamma\alpha_2\beta + \beta^2). \end{aligned}$$

Excepting all the cases where $x^* = 0$ or where $y^* = 0$, this implies ℓ' has two zeros according to the quadratic formula. Because $\alpha_1 > \alpha_2$ we have $\ell'(0) > 0$, and thus ℓ is increasing at $w = 0$. This implies that if the smaller of the zeros of $f(w)$ is greater than one, then $\hat{w} \equiv 1$. Using the quadratic formula with

$$a = -\gamma^2\alpha_1 - \gamma^2\alpha_2 - \gamma^2,$$

$$b = \gamma\alpha_1\beta - \gamma^2\alpha_1 - \gamma\alpha_2\beta - \gamma^2\alpha_2,$$

$$c = \gamma\alpha_1\beta - \gamma\alpha_2\beta + \beta^2,$$

the smaller zero is

$$\hat{w} = \frac{-b - \sqrt{b^2 - 4ac}}{2a},$$

which is a function of the generating parameters x^* and y^*. We see that

$a \leq 0$ and, by a small calculation, $2a + b \leq 0$. With this we have,

$$\hat{w} \geq 1 \iff |2a + b| \leq \sqrt{b^2 - 4ac} \iff a + b + c \geq 0.$$

Using

$$\alpha_1 + \alpha_2 = \frac{1}{4}\beta$$

and

$$\alpha_1 - \alpha_2 = x^*(y^*)^2,$$

and simplifying as functions of γ and β shows that $a+b+c \geq 0$ is equivalent to

$$-\gamma^2\left(1 + \frac{1}{2}\beta\right) + 2\gamma\beta x^*(y^*)^2 + \beta^2 \geq 0. \tag{16}$$ {eq:restricted-bl-re

□

Theorem 3. *Define* $\gamma(x, y) = 4xy$. *For*

$$\beta := \beta(x^*, y^*) = 1 + (x^*)^2 + (y^*)^2 + (x^*)^2(y^*)^2,$$

$$\gamma := \gamma(x^*, y^*),$$

bounds

$$\gamma_L := \gamma_L(x^*, y^*) \leq \gamma(\hat{x}, \hat{y}),$$

$$\gamma_U := \gamma_U(x^*, y^*) \geq \gamma(\hat{x}, \hat{y}),$$

and

$$\beta_L := \beta_L(x^*, y^*) \leq \beta(\hat{x}, \hat{y}),$$

the maximum likelihood value $\hat{w} = 1$ *when*

$$-\gamma_U^2\left(1 + \frac{1}{2}\beta\right) + 2\gamma_L\beta_L x^*(y^*)^2 + \beta_L^2 \geq 0.$$

Proof. In the general case, $\hat{w}$ is a function of x^*, y^*, $\hat{x}$, and $\hat{y}$. From the previous section, $\hat{w}$ is given by the quadratic formula, though now with γ and β as functions of $\hat{x}$ and $\hat{y}$ instead of x^* and y^*. Assume we know $\hat{x}$ and $\hat{y}$ as functions of x^* and y^* only. The same derivation as for (16) further

yields

$$-\gamma^2(\hat{x},\hat{y})\left(1+\frac{1}{2}\beta\right)+2\gamma(\hat{x},\hat{y})\beta(\hat{x},\hat{y})x^*(y^*)^2+\beta^2(\hat{x},\hat{y})\geq 0 \qquad (17) \quad \{\texttt{eq:general-bl-resul}$$

where, since α_1 and α_2 are still functions of x^* and y^*, $(1+1/2\cdot\beta)$ and $x^*(y^*)^2$ from (16) remain unchanged. Given the bounds

$$\gamma_L \leq \gamma(\hat{x},\hat{y}) \leq \gamma_U$$

and

$$\beta_L \leq \beta(\hat{x},\hat{y}),$$

if

$$-\gamma_U^2\left(1+\frac{1}{2}\beta\right)+2\gamma_L\beta_L x^*(y^*)^2+\beta_L^2\geq 0,$$

then (17) holds, and this is an inequality involving only x^* and y^*. □

To get a sense of the region of overestimation, Fig. 4 plots an intermediate case.

Empirical validation

We use basin-hopping [Wales and Doye, 1997] implemented in `scipy` [Jones et al., 2001–] to obtain the plots in Figs. 5, S2, and S3. This method randomly perturbs candidate solutions and accepts or rejects them based on the nearby likelihood surface, eventually obtaining an estimate of the "global" optimum. While it is not guaranteed to always obtain a global optimum—or even always converge—its implementation here leads us to believe that we can estimate $\hat{w}$ as well as any optimization method available.

We take various safeguards to ensure convergence and stability of the procedure. Since this method involves randomness, we take precautions to not evaluate the likelihood near or outside of boundary conditions, i.e., when x^* or y^* are either 0 or 1. As is apparent from Table S1, if $0<x^*,y^*<1$ and we try to evaluate the candidate $\hat{x}=\hat{y}=\hat{w}=1$ we have a likelihood of exactly zero. Worse still, in Table S4 for the Farris tree, in the cases

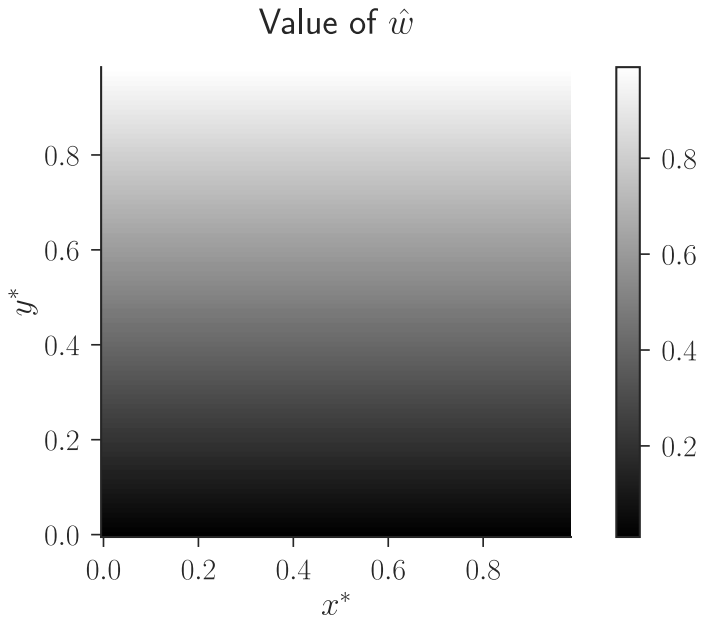


Figure S2: Estimates for $\hat{w}$ when computing $(\hat{x}, \hat{y}, \hat{w})$ using basin-hopping [Wales and Doye, 1997] optimizing the classical marginal likelihood (2) (rather than a joint optimization procedure). {fig:bl-general-marg

of $\{2\}, \{1,2\}, \{2,3\}$, and $\{1,2,3\}$, any attempt to evaluate $\hat{y} = 1$ will similarly yield a zero likelihood. In these cases of zero likelihood, methods that evaluate many candidate parameter values can fail to converge when near these boundaries. We sidestep these computational issues by restricting ourselves to the region $x^*, y^* \in [10^{-2}, 1 - 10^{-2}]^2$.

We initialize our optimization procedure at the true branch parameters $t^* = \{x^*, y^*, x^*, y^*, y^*\}$. Because our analysis shows that $\hat{w}$ can be equal to one when x^* and y^* are small, and local optimization may have a hard time obtaining this value if our initial guess is small, we perform two maximizations, one with $w = 1$ fixed and one where w is estimated. We take the value of $\hat{w}$ with the larger objective function as our estimate.

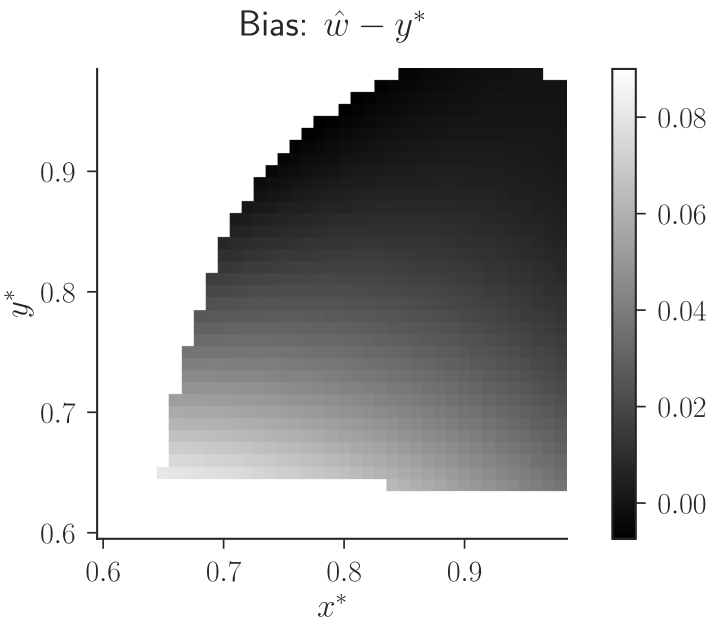


Figure S3: Estimates for $\hat{w} - y^*$ when computing $(\hat{x}, \hat{y}, \hat{w})$ using basin-hopping [Wales and Doye, 1997] optimizing (3). Plot focuses on $0.6 < x^*, y^* < 1$ where $\hat{w}$ is estimated to be a value different from 1. We do not compute the bias for the white region where $\hat{w} = 1$ to preserve a useful scale for the rest of the plot.

{fig:bl-general-bias