Published in final edited form as:

Proc IEEE Int Symp Info Theory. 2016 July ; 2016: 580–584. doi:10.1109/ISIT.2016.7541365.

Chained Kullback-Leibler Divergences

Dmitri S. Pavlichin and
Stanford University

Tsachy Weissman
Stanford University

Abstract

We define and characterize the “chained” Kullback-Leibler divergence $\min_w D(p\|w) + D(w\|q)$ minimized over all intermediate distributions w and the analogous k -fold chained K-L divergence $\min D(p\|w_{k-1}) + \dots + D(w_2\|w_1) + D(w_1\|q)$ minimized over the entire path $(w_1, \dots, w_{k-1})$. This quantity arises in a large deviations analysis of a Markov chain on the set of types – the Wright-Fisher model of neutral genetic drift: a population with allele distribution q produces offspring with allele distribution w , which then produce offspring with allele distribution p , and so on.

The chained divergences enjoy some of the same properties as the K-L divergence (like joint convexity in the arguments) and appear in k -step versions of some of the same settings as the K-L divergence (like information projections and a conditional limit theorem). We further characterize the optimal k -step “path” of distributions appearing in the definition and apply our findings in a large deviations analysis of the Wright-Fisher process. We make a connection to information geometry via the previously studied continuum limit, where the number of steps tends to infinity, and the limiting path is a geodesic in the Fisher information metric.

Finally, we offer a thermodynamic interpretation of the chained divergence (as the rate of operation of an appropriately defined Maxwell’s demon) and we state some natural extensions and applications (a k -step mutual information and k -step maximum likelihood inference). We release code for computing the objects we study.

I. Introduction

We investigate the properties and applications of the k -fold chained Kullback-Leibler divergence between distributions p and q :

$$D^{(k)}(p\|q) \triangleq \min_{w_1, \dots, w_{k-1}} \sum_{i=1}^k D(w_i\|w_{i-1}) \quad (1)$$

where $w_0 \triangleq q$, $w_k \triangleq p$, $D(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence

$$D^{(k)}(p||q) = \sum_{x \in \mathcal{X}} p[x] \log \frac{p[x]}{q[x]} \quad (2)$$

and the minimum, if it exists, is taken over a “path” of distributions w_i in some family of distributions. k counts the number of hops and $D^{(1)}(p||q) \triangleq D(p||q)$.

Just as the K-L divergence appears in statistical applications of information theory and in the method of types [1], so the chained divergence arises in the analysis of stochastic processes with an iterative resampling flavor: in each round one samples from a distribution w_i determined by the outcome of the previous round of sampling from w_{i-1} . A motivating example that we consider in detail is the Wright-Fisher model of neutral genetic drift [2], [3], wherein each generation is obtained by sampling with replacement from the previous generation.

This work is devoted to characterizing the chained divergences in the finite alphabet setting and the optimal k -step path appearing in their definition, and to pointing out their applications. Section II defines the chained divergence in tandem with a discussion of the Wright-Fisher process, in which it naturally appears. Section III derives some properties (like joint convexity) of the chained divergence, including several “operationally” meaningful properties characterizing the rate of large deviations, information projections, and a conditional limit theorem; these are very similar to results on the K-L divergence [4], [5], [6]. Section IV-A characterizes the optimal path of intermediate distributions ($w_1, \dots, w_{k-1}$) for all values of k and states a method for computing this path. Section IV-B considers the continuum limit of $k \rightarrow \infty$ (letting the number of distributions in the optimal path tend to infinity) and recapitulates results previously obtained by [7] in this setting: the limiting path is a geodesic in the Fisher information metric; equivalently, part of a great circle connecting $\sqrt{p}$ and $\sqrt{q}$ ¹. Section V considers extensions and other applications of the chained divergences – in particular a k -step version of the mutual information related to the likelihood that two independent genetic loci become dependent through neutral genetic drift in k generations. Finally, we offer a thermodynamic interpretation of the chained divergence as the optimal rate of operation of an appropriately defined Maxwell’s demon.

We release code for computing the objects we study at². An Appendix contains the details of some of the proofs omitted here due to space constraints and is available at [8].

II. The Wright-Fisher Markov chain

We define the Wright-Fisher (“iterative resampling”) process and review briefly some results about the K-L divergence that we extend to the chained divergences (1).

¹ $\sqrt{p}$ denotes the vector with components $\sqrt{p[x]}$.

² <https://github.com/dmitrip/chaineddivergences>

Denote by $\Delta \parallel \left\{ p \in \mathbb{R}^{\mathcal{X}} : \sum_{x \in \mathcal{X}} p[x] = 1, p \geq 0 \right\}$ the set of distributions on finite alphabet $\mathcal{X}$. Denote by Δ_n the set of distributions with integer denominator n for each component:

$$\Delta_n \parallel \Delta \cap \left\{ p : np \in \mathbb{N}^{\mathcal{X}} \right\} \quad (3)$$

Δ_n is the set of types [1], equivalently the possible empirical distributions of n $\mathcal{X}$ -valued samples.

Consider the Markov chain whose state space is the set of types Δ_n (3) defined by the following rule: Let the current state of the chain be $q \in \Delta_n$. Draw n samples x_1^n i.i.d. from q

and let $p[x] = \frac{1}{n} |\{i : x_i = x\}|$ be the empirical distribution. p is the next state of the chain. This is the Wright-Fisher model [2], [3] of neutral genetic drift among $|\mathcal{X}|$ alleles in a population of size n : Each of n individuals in a generation is sampled by cloning a uniformly chosen (hence “neutral”) individual from the previous generation. The process is illustrated in Figure 1. As time increases, alleles “die out” (once $q[x] = 0$, no samples of x can be observed again) until only one allele remains, or “fixes” (that is, $q[x] = 1$ for some $x \in \mathcal{X}$), so the vertices of the simplex Δ_n are absorbing states. We will study fluctuations of this chain before fixation.

Denote by $P_n(p|q)$ the transition matrix for this Markov chain:

$$P_n(p|q) = \text{MultinomialPMF}(np; q) \quad (4)$$

$$= \binom{n}{np[x]_{x \in \mathcal{X}}} \prod_{x \in \mathcal{X}} q[x]^{np[x]} \quad (5)$$

$$\doteq_n e^{-nD(p||q)} \quad (6)$$

where the prefactor in (5) is a multinomial coefficient, $\doteq_n$ denotes equality to leading exponential order in n^3 , and $D(\cdot||\cdot)$ denotes the Kullback-Leibler divergence (2). We can evaluate $P_n(p|q)$ for any $p, q \in \Delta_n$ and n , but since the state space for the Wright-Fisher process is the set of types Δ_n (3), $n \in \mathbb{N}$, we can think of $P_n(p|q)$ as a $|\Delta_n| \times |\Delta_n|$ stochastic matrix.

³That is, $a_n \doteq_n b_n \iff \lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$.

Given a set of distributions $E \subset \Delta_n$, we define $P_n(E|q) \parallel \sum_{p \in E \cap \Delta_n} P_n(p|q)$ as the probability that the chain hops to somewhere in E starting from q . If $E \cap \Delta_n = \emptyset$, then $P_n(E|q) \triangleq 0$. If E is the closure of its interior (so $E \cap \Delta_n$ is nonempty for all large enough n) then Sanov's theorem tells us [4], [5]

$$P_n(E|q) \parallel \sum_{p \in E \cap \Delta_n} P_n(p|q) \doteq e^{-nD(E||q)} \quad (7)$$

where

$$D(E||q) \parallel \inf_{p \in E} D(p||q) = D(p^*||q) \quad (8)$$

where $p^* \triangleq \arg \min_{p \in E} D(p||q)$ is the I-projection [6], where the minimum exists and is unique since E is closed (thus compact since $E \subset \Delta_n$) by assumption and $D(p||q)$ is strictly convex in p . Sanov's theorem and the conditional limit theorem tell us [4], [5], [1] that conditioned on drawing empirical distribution $p \in E$, distribution p is close⁴ to p^* as $n \rightarrow \infty$. If E is not closed (so $\min_{p \in E} D(p||q)$ might not exist) but is convex, then there exists a unique distribution p^* – the generalized I-projection [5] – such that $D(p||q) \geq D(p||p^*) + D(E||q)$ for all $p \in E$.

What happens when we iterate the resampling chain? Denote by $P_n^{(k)}(p|q)$ the probability to draw distribution p starting from q in k steps, corresponding to the transition matrix

$P_n^{(k)} = P_n^k$. We recursively express $P_n^{(k)}(\cdot|q)$:

$$P_n^{(k)}(p|q) \parallel \sum_{w \in \Delta_n} P_n(p|w) P_n^{(k-1)}(w|q) \quad (9)$$

$$P_n^{(k)}(E|q) \parallel \sum_{w \in \Delta_n} P_n(E|w) P_n^{(k-1)}(w|q) \quad (10)$$

with $P_n^{(1)}(\cdot|q) \parallel P_n(\cdot|q)$ as in (4) and (7).

Theorem 1 establishes that the k -fold chained K-L divergence (1) plays the same role for the k -step resampling chain as the K-L divergence plays for one step of the chain, as described above. It is convenient to define recursively the k -step chained divergence $D^{(k)}(p||q)$ for $p, q \in \Delta_n$:

⁴Convergence in probability, as in the statement of Theorem 1.6) below.

$$D^{(k)}(p\|q) \triangleq \min_{w \in \Delta} \left(D(p\|w) + D^{(k-1)}(w\|q) \right) \quad (11)$$

with $D^{(1)}(p\|q) \triangleq D(p\|q)$, the K-L divergence (2). Note that we optimize over the simplex (rather than the types $\mathcal{T}_n$). Theorem 1 establishes the existence and uniqueness of the minimum and the equivalence of the recursive definition (11) with our earlier definition for the chained divergence (1).

We further define $D^{(k)}(E\|q)$ for a closed, convex set $E \subset \Delta$ by analogy with (8):

$$D^{(k)}(E\|q) \triangleq \inf_{p \in E} D^{(k)}(p\|q) = D^{(k)}(p^{(k)*}\|q) \quad (12)$$

and $p^{(k)*} \triangleq \arg \min_{p \in E} D^{(k)}(p\|q)$ is the I-projection of $w_{k-1}^{(k)*}$ on E , where $w_{k-1}^{(k)*}$ is the next-to-last point in the optimal path (15). If E is convex but $\min_{p \in E} D^{(k)}(p\|q)$ does not exist, then there exists a unique distribution $p^{(k)*}$ that is the generalized I-projection of $w_{k-1}^{(k)*}$ on E .

III. Characterizing the k -fold chained K-L divergence

Theorem 1

Let k be a positive integer.

1. The k -fold chained divergence (11) satisfies for all $p, q \in \Delta$ such that $D(p\|q) < \infty$:

$$D^{(k)}(p\|q) \triangleq \min_{w \in \Delta} \left(D(p\|w) + D^{(k-1)}(w\|q) \right) \quad (13)$$

$$= \sum_{i=1}^k D(w_i^{(k)*}\|w_{i-1}^{(k)*}) \quad (14)$$

with boundary conditions $w_k^{(k)*}\|p, w_0^{(k)*}\|q$, and

$$(w_1^{(k)*}, \dots, w_{k-1}^{(k)*}) \triangleq \arg \min_{\Delta^{k-1}} \sum_{i=1}^k D(w_i\|w_{i-1}) \quad (15)$$

where the minimum is over the $(k-1)$ -fold product set $\Delta^{k-1} = \Delta \times \dots \times \Delta$. The minimum exists and is unique. For closed, convex $E \subset \Delta$, (12) defines $D^{(k)}(E\|q)$.

If $D(p\|q) = \infty$, then $D^{(k)}(E\|q) \triangleq \infty$ and $w_i^{(k)*}$ is not defined.

2. Convexity: $D^{(k)}(E||q)$ is jointly strictly convex in p and q :

$$D^{(k)}(\alpha p_1 + (1 - \alpha) p_2 || \alpha q_1 + (1 - \alpha) q_2) < \alpha D^{(k)}(p_1 || q_1) + (1 - \alpha) D^{(k)}(p_2 || q_2) \quad (16)$$

for all distributions $(p_1, q_1), (p_2, q_2), \alpha \in (0, 1)$.

3. Scaling in k : Suppose $0 < D(p||q) < \infty$. Then $D^{(k)}(E||q)$ is strictly decreasing in k and $D^{(k)}(p||q) = O\left(\frac{1}{k}\right)$. Theorem 3 of Section IV-B (on the continuum limit $k \rightarrow \infty$) gives the more precise scaling:

$$\lim_{k \rightarrow \infty} k D^{(k)}(p||q) = 2\theta^2_{\sqrt{p}, \sqrt{q}} \quad (17)$$

where $\theta_{\sqrt{p}, \sqrt{q}} = \arccos(\sqrt{p} \cdot \sqrt{q})$ is the angle between vectors $\sqrt{p}$ and $\sqrt{q}$ (see Section IV-B for details).

4. Markov chain transition matrix $P_n^{(k)}(p|q)$ (9), upper bound for $p \in \mathcal{X}$:

$$P_n^{(k)}(p|q) \leq (n+1)^{(k-1)|\mathcal{X}|} e^{-nD^{(k)}(p||q)} \quad (18)$$

Moreover, for $p \in \mathcal{X}$ and sequence $(p_n)_{n \in \mathbb{N}}, p_n \in \mathcal{X}$ such that $\lim_{n \rightarrow \infty} D(p_n||p) = 0$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P_n^{(k)}(p|q) = \lim_{n \rightarrow \infty} \frac{1}{n} \log P_n^{(k)}(p_n|q) = -D^{(k)}(p||q) \quad (19)$$

Equivalently, $P_n^{(k)}(p|q) \doteq e^{-nD^{(k)}(p||q)}$.

5. Sanov's theorem: Let $E \subset \mathcal{X}$. The probability $P_n^{(k)}(E|q)$ (10) to draw a type in E from $q \in \mathcal{P}(\mathcal{X})$ in k steps is upper bounded:

$$P_n^{(k)}(E|q) \leq (n+1)^{k|\mathcal{X}|} e^{-nD^{(k)}(E||q)} \quad (20)$$

where $D^{(k)}(E||q)$ is defined in (12). Moreover, if E is closed, convex, and $E \cap \mathcal{P}_n \neq \emptyset$ for sufficiently large n then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P_n^{(k)}(E|q) = -D^{(k)}(p^{(k)*}||q) \quad (21)$$

where $p^{(k)*}$ the k -step I-projection of q on E (12); equivalently,

$$P_n^{(k)}(E|q) \doteq e^{-nD^{(k)}(E||q)}.$$

6. Conditional limit theorem: Let $E \subset \mathcal{X}$ be closed, convex, and $E \cap \text{Support}(q) \neq \emptyset$ for sufficiently large n , $q \in \mathcal{P}(\mathcal{X})$, $D(E||q) < \infty$. Let $(X_t^{(i)})_{t=1}^n$ be the i -th sample of the Wright-Fisher process (Figure 1): that is, $X_t^{(i)}$ drawn i.i.d. $\sim \hat{w}_{i-1}$ with $\hat{w}_0 = q$ and $\hat{w}_i$ the empirical distribution of $X_t^{(i)}$. Then for all $\varepsilon > 0$, $i \in \{1, \dots, k\}$, $x \in \mathcal{X}$

$$\lim_{n \rightarrow \infty} (|\hat{w}_i[x] - w_i^*[x]| \geq \varepsilon | \hat{w}_k \in E) = 0 \quad (22)$$

with $(w_i^{(k)*})_{i=1}^{k-1}$ as in (15) and $w_k^* = p^{(k)*} = \arg \min_{p \in E} D^{(k)}(p||q)$ (the k -step I-projection of q on E (12)).

Joint convexity 2) follows because minimization of jointly convex functions with respect to some of the arguments over a convex set preserves convexity [9]. Existence and uniqueness of the optimal path 1) follow from joint strict convexity and the compactness of $\cap \text{Support}(q)$. Theorem 3 contains 3). The proofs of 4), 5), 6) all follow closely the proofs of analogous results for the K-L divergence in [1], Chapter 11. See Appendix VI-A in [8] for details of the proofs.

IV. Characterizing the k -fold path

What can we say about the optimal k -fold “path” $(w_i^{(k)*})_{i=1}^{k-1} = \min_{\Delta^{k-1}} \sum_{i=1}^k D(w_i||w_{i-1})$ (15)? In the event of a large deviation, we observe a path close to this one with high probability as $n \rightarrow \infty$ (Theorem 1.6)).

A. Finite number of steps k

Let’s start with the case $k = 2$ and find $w^*(p, q) = \arg \min_{w \in \mathcal{X}} D(p||w) + D(w||q)$. For $k > 2$, we will then use this local characterization of the path of distributions to compute

$$w_i^* = w^*(w_{i+1}^*, w_{i-1}^*).$$

We set up an optimization problem with Lagrangian

$\Lambda = D(p||w) + D(w||q) + (\lambda - 1) \left(\sum_{x \in \mathcal{X}} w[x] - 1 \right)$ where $\lambda - 1$ is a Lagrange multiplier⁵ enforcing normalization of w . Solving the $|\mathcal{X}|$ equations $\nabla \Lambda|_{w^*} = 0$ for w^* we find:

$$w^*[x] = w^*(p, q)[x] \parallel \begin{cases} \frac{p[x]}{W(e^{\lambda^* \frac{p[x]}{q[x]}})} & : p[x] > 0 \\ e^{-\lambda^*} q[x] & : p[x] = 0 \end{cases} \quad (23)$$

⁵We use $\lambda - 1$ rather than λ to make the answer look more compact.

where $W(\cdot)$ is the principal branch of the Lambert W function⁶ and λ^* is chosen so that $w^* \in \mathcal{X}$ is normalized (in fact $\lambda^* \in [0, 1]$ for all p, q). The case of $p[x] = 0$ is obtained by analytic continuation of the solution of the case $p[x] > 0$. if $D(p||q) = \infty$ then $w^*(p, q)$ is not defined. Note that the Lagrange multiplier λ^* lives inside the Lambert W function (rather than outside like a partition function prefactor), and so can not generally be expressed analytically in terms of p and q ; one could try to find λ^* numerically⁷. Theorem 1.1) tells us that w^* is a global minimum.

Theorem 2—(local characterization of the optimal k -step path) Let $k \geq 2$ and $D(p||q) < \infty$ for $p, q \in \mathcal{P}$. The optimal path $\left(w_i^{(k)*}\right)_{i=1}^{k-1}$ (15) used in defining the chained K-L divergence $D^{(k)}(p||q)$ satisfies

$$w_i^{(k)*}[x] = w^*\left(w_{i+1}^{(k)*}, w_{i-1}^{(k)*}\right)[x] \quad (24)$$

$$\geq p[x]^{i/k} q[x]^{(k-i)/k} \quad (25)$$

for all $x \in \mathcal{X}$, $1 \leq i \leq k-1$, where $w^*(\cdot, \cdot)$ is defined in (23) and with boundary conditions $w_k^{(k)*} = p, w_0^{(k)*} = q$. As a corollary of (24), for $i \in \{1, \dots, k-1\}$

$$\text{Support}\left(w_i^{(k)*}\right) = \text{Support}(q) \quad (26)$$

Proof—Suppose (24) does not hold for some $p, q, w_i^{(k)*}$. Then $\sum_{i=1}^k D\left(w_i^{(k)*} || w_{i-1}^{(k)*}\right)$ (14) can be strictly decreased by replacing $w_i^{(k)*}$ with $w^*\left(w_{i+1}^{(k)*}, w_{i-1}^{(k)*}\right)$. Thus (24) holds. ■

To derive the geometric lower bound (25) we massage the function $w^*(p, q)$ (23) into an f -divergence⁸, then use the non-negativity of f -divergences to bound λ^* and derive $w^*(p, q)[x] \geq \sqrt{p[x]q[x]}$, and then derive (25). See Appendix VI-B [8] for details.

Remarks: Figure 2 plots the optimal k -step path for $k \in \{2, 6\}$ and in the $k \rightarrow \infty$ limit (see Section IV-B). One way to compute the optimum path is to start with some guess (a good initial point is given in Theorem 3) and repeatedly apply (24) until numerical convergence; this is the method used in making Figure 2 and implemented in the code we release⁹.

⁶ $W(z)$ satisfies $z = W(z)e^{W(z)}$ and $W(z) \geq -1$ corresponds to the principal branch.

⁷ This means finding a root of the smooth, monotonic function of λ $\sum_{x \in \mathcal{X}} p[x]/W\left(e^{\lambda} \frac{p[x]}{q[x]}\right) - 1$ with the knowledge that $\lambda^* \in [0, 1]$, so is doable.

⁸ The f -divergence is $D_f(p, q) = \sum_{x \in \mathcal{X}} q[x] f(p[x]/q[x])$ with $f(t) = -t' W(te^\lambda) + t' W(e^\lambda)$. $f(t)$ is convex and $f(1) = 0$, so $D_f(p, q)$ is an f -divergence.

The optimal k -step path from q to p is not the same as the optimal path from p to q – an asymmetry inherited from the asymmetry of $D^{(k)}(p||q)$; but in the limit $k \rightarrow \infty$ these two paths converge to the same limiting path (see Section IV-B). From (23) we see that $q[x] > 0 \Rightarrow w^*(p, q)[x] > 0$, so $\text{Support}(w_i^{(k)*}) = \text{Support}(q)$ for $i \in \{1, \dots, k-1\}$.

B. The limit $k \rightarrow \infty$

What happens in the continuum limit, as the number of steps $k \rightarrow \infty$? This setting was investigated by [7], who states the limiting path and large deviations rate function. We include this Section to make the story more complete, and to check consistency with and provide a finite- k intuition for the results [7] obtained with variational calculus.

A useful perspective in the following is to map the simplex to the part of the unit sphere in

$\mathbb{R}^{|\mathcal{X}|}$ in the non-negative orthant, $\Psi = \{\psi \in \mathbb{R}^{|\mathcal{X}|} : \|\psi\|_2 = 1, \psi \geq 0\}$ via the bijection

$p \rightarrow \sqrt{p} \parallel \left(\sqrt{p[x]} \right)_{x \in \mathcal{X}}$. This square root reparametrization appears in [10], [11], [12], [13].

Consider the K-L divergence between nearby points on the orthant Ψ . That is, if $p = q + \varepsilon \in \Psi$,

$\varepsilon \in \mathbb{R}^{|\mathcal{X}|}$, then $\exists \delta_\varepsilon \in \mathbb{R}^{|\mathcal{X}|}$ such that $\sqrt{p} = \sqrt{q} + \delta_\varepsilon \in \Psi$ where

$\delta_\varepsilon[x] = \frac{1}{2} \varepsilon[x] p[x]^{-1/2} + O(\varepsilon[x]^2 p[x]^{-3/2})$. Then

$$D(p||q) = 4 \left(1 - \overbrace{\sqrt{p} \cdot \sqrt{q}}^{\cos \theta} \right) + \sum_{x \in \mathcal{X}} O \left(\frac{\delta_\varepsilon[x]^3}{\sqrt{q[x]}} \right) \quad (27)$$

$$= 2\theta^2 + O(\theta^4) + \sum_{x \in \mathcal{X}} O \left(\frac{\theta^3}{\sqrt{q[x]}} \right) \quad (28)$$

where in (27) we rewrote the higher order terms in terms of ε , expanded the cos about $\theta = 0$, and where

$$\theta = \theta_{\sqrt{p}, \sqrt{q}} = \arccos(\sqrt{p} \cdot \sqrt{q}) \quad (29)$$

is the angle between vectors $\sqrt{p}$ and $\sqrt{q}$ and $\sqrt{p} \cdot \sqrt{q} = B(p, q)$ is the Bhattacharyya coefficient between p and q .

⁹<https://github.com/dmitrip/chaineddivergences>

The above recapitulates the familiar fact that the K-L divergence is close to the squared Euclidean distance between nearby distributions [13] (here written in square root space); the leading order term is symmetric in p and q and depends on p and q only through the angle $\theta_{\sqrt{p}, \sqrt{q}}$. This enables us to guess that the limiting optimal path is a geodesic in the Euclidean metric restricted to the surface of the unit sphere and that the angle between adjacent points

$\theta_{\sqrt{w_i^{(k)*}}, \sqrt{w_{i-1}^{(k)*}}} \rightarrow \frac{1}{k} \theta_{\sqrt{p}, \sqrt{q}}$. Theorem 3 confirms this intuition. See Appendix VI-C [8] for proof details.

Theorem 3—Let $D(p||q)$, $D^{(k)}(p||q)$ be as in (2), (14), respectively. Let $p, q \in \mathcal{P}$ and $D(p||q) < \infty$ and $D(q||p) < \infty$.

1.

Scaling of $D^{(k)}(p||q)$ in k : $D^{(k)}(p||q) = O\left(\frac{1}{k}\right)$ and

$$\lim_{k \rightarrow \infty} k D^{(k)}(p||q) = \lim_{k \rightarrow \infty} k D^{(k)}(q||p) = 2\theta_{\sqrt{p}, \sqrt{q}}^2 \quad (30)$$

where (as in (29)) $\cos \theta_{\sqrt{p}, \sqrt{q}} = \sqrt{p} \cdot \sqrt{q} = B(p, q)$.

2.

The limiting path $(w_i^{(k)*})$ (15) as $k \rightarrow \infty$ is a geodesic in the Euclidean metric restricted to the unit sphere (equivalently, a geodesic in the Fisher information metric on the simplex): part of a great circle connecting $\sqrt{p}$ and $\sqrt{q}$ with constant angular speed. Let $t \in [0, 1]$, then

$$w_{\lfloor tk \rfloor}^{(k)*} \rightarrow \frac{1}{Z_{\tau_t}} (\tau_t \sqrt{p} + (1 - \tau_t) \sqrt{q})^2 \quad (31)$$

in L^2 norm as $k \rightarrow \infty$, where $\lfloor x \rfloor$ denotes the floor function, Z_{τ_t} is a constant¹⁰, and $\tau_t: [0, 1] \rightarrow [0, 1]$ is a reparametrization of “time” t that ensures constant angular speed¹¹ on the unit sphere.

Remarks: The condition $D(p||q) < \infty$ and $D(q||p) < \infty$ ensures that $p[x] > 0 \Leftrightarrow q[x] > 0$, so the series expansion (27) does not diverge and the limits in (30) match. Figure 2 depicts the limiting optimal path for a particular choice of p and q . For a finite number of steps k , the limiting value of the quantity $w_{\tau_i/k}^{(k)*}$ (31) provides a decent initial guess for the iterative computation of $w_i^{(k)*}$ described in the remarks following the statement of Theorem 2.

¹⁰We can check that $Z_{\tau_t} = \tau_t^2 + (1 - \tau_t)^2 + 2\tau_t(1 - \tau_t)(\sqrt{p} \cdot \sqrt{q})$.

¹¹We can check to $\tau_t = \frac{1}{2} \left(1 + \frac{\tan((t - 1/2)\theta)}{\tan(\theta/2)} \right)$ where $\theta = \theta_{\sqrt{p}, \sqrt{q}}$.

V. Applications and further directions

We conclude by offering an application of the chained K-L divergences in maximum likelihood inference and an interpretation in terms of a thought experiment in thermodynamics.

A. ML inference and mutual information

The k -step K-L divergence $D^{(k)}(p\|E) = \inf_{q \in E} D^{(k)}(p\|q)$ “from” a set of distributions E has a maximum likelihood interpretation: $q^* = \min_{q \in E} D^{(k)}(p\|q)$ maximizes the likelihood to draw empirical distribution $p \in \mathcal{P}$ in k steps of the Wright-Fisher process with initial distribution $q \in E$. A special case of this is a k -step generalization of the mutual information $I(X; Y)$ between random variables X and Y with finite alphabets $\mathcal{X}, \mathcal{Y}$, respectively, jointly distributed as $p_{XY} \in \Delta^{\mathcal{X} \times \mathcal{Y}}$:

$$I^{(k)}(X; Y) = \min_{q_X, q_Y} D^{(k)}(p_{XY} \| q_X q_Y) \quad (32)$$

with the minimum attained by the k -step “marginals”

$(q_X^{(k)*}, q_Y^{(k)*}) = \operatorname{argmin}_{q_X, q_Y} D^{(k)}(p_{XY} \| q_X q_Y)$. We can check that the minimum exists and is unique¹². For $k = 1$, we have $I^{(1)}(X; Y) = I(X; Y)$ and $(q_X^{(1)*}, q_Y^{(1)*}) = (p_X, p_Y)$ (the marginals of p_{XY}), but for $k > 1$ this is not the case. $I^{(k)}$ is monotonically decreasing in k .

Written in the minimization form (32), computing the mutual information $I(X; Y)$ corresponds to finding the maximum likelihood distribution under which X and Y are independent given data with empirical distribution p_{XY} . In the context of the Wright-Fisher process, suppose we assume that two genetic loci X and Y are independently distributed k generations ago, but are no longer independent due to the intervening neutral genetic drift; then the maximum likelihood ancestral distribution is $q_X^{(k)*} q_Y^{(k)*}$ given the current distribution p_{XY} . After enough time, all but one pair (x, y) of alleles fixes and the two loci become independent again, but neutral drift induces dependence before fixation occurs.

A related object we can construct with the chained K-L divergence is $D^{(k)}(p_{XY} \| p_X p_Y)$; this, like (32), matches the mutual information $I(X; Y)$ for $k = 1$. The authors do not yet give an operational meaning for this quantity.

B. Maxwell’s demon

Maxwell’s demon is a thought experiment in thermodynamics, here envisioned as a model of a desalination plant: let distribution q correspond to the ambient relative concentrations of some chemicals in the “ocean” (so $\mathcal{X} = \{\text{sodium ion, water, potassium ion, } \dots\}$). The demon’s goal is to achieve a desired concentration of chemicals in the water supply; let E

¹²The set of distributions $\{q_{XY} : q_{XY} = q_X q_Y\}$ is log-convex, and letting $q_{XY}^{(k)*}$ be the “reverse I-projection” of $w_{k-1}^{(k)*}$ (15) on E , Theorem 1 of [6] proves existence and uniqueness.

denote the set of concentrations the demon considers acceptable. The demon operates by admitting n molecules drawn from the ambient distribution q into a vesicle (temporary storage); if it so happens that p – the empirical distribution of the n molecules – satisfies $p \in E$, then the demon releases the contents of the vesicle into the water supply, and otherwise the demon dumps the contents back in the ocean. See Figure 3.

How quickly does the demon desalinate? From the preceding discussion we know that the demon releases the n molecules into the water supply with probability about $e^{-nD(E||q)}$. Suppose we now break up the desalination process into two demons, so there is an intermediate stage between the ocean and the water supply. If either of the two demons rejects, then both reject¹³. Then the optimal 2-demon desalination rate is $e^{-nD^{(2)}(E||q)}$ ¹⁴, which is about twice as large as the 1-demon rate. Adding more stages increases the rate further; in the $k \rightarrow \infty$ limit, we have a concentration gradient across the k stages with no large “jumps.” This is not how actual desalinators work, but the design principle of breaking up a large concentration difference into many small stages is used in multi-stage flash desalination and in the loop of Henle in the kidney.

Acknowledgments

The authors gratefully acknowledge Surya Ganguli and Hideo Mabuchi for suggestions and insightful discussions.

VI. Appendix

A. Proof of Theorem 1

1) (Equivalence of recursive and non-recursive definitions of the chained divergence)

We proceed by induction on k . For the base case $k = 2$ the two definitions (13) and (14) are the same. Now suppose the definitions are equivalent up to $k - 1$ (abusing notation). Then we expand out the min:

$$D^{(k)}(p||q) = \min_{w_{k-1}} \left(D(p||w_{k-1}) + D^{(k-1)}(w_{k-1}||q) \right) \quad (33)$$

$$= \min_{w_{k-1}} \left(D(p||w_{k-1}) + \min_{w_1, \dots, w_{k-2}} \sum_{i=1}^{k-1} D(w_i||w_{i-1}) \right) \quad (34)$$

¹³This ensures particle conservation in the intermediate stage, though we can envision other scenarios with variable sample sizes and demon policies.

¹⁴The intermediate demon must choose an appropriate acceptance set $E' \subset E$ such that the I-projection of q on E' is $w^* = \arg \min_w D(E||w) + D(w||q)$; one possible choice is $E' = \{w: D(E||w) = D(E||w^*)\}$.

$$\min_{w_1, \dots, w_{k-1}} \sum_{i=1}^k D(w_i \| w_{i-1}) \quad (35)$$

where the second line follows from the inductive hypothesis. $\square$

Existence and uniqueness of the path follows from the joint strict convexity of $D^{(k)}(p \| q)$ (Theorem 1.2)).

2) (Joint convexity of $D^{(k)}(\cdot \| \cdot)$)

We use the following result [9] (Section 3.2.5): Suppose function $f(x, y, z)$ is jointly convex in $x, y, z \in C$, where C is a nonempty convex subset of $\mathbb{R}^d$. Then

$$g(x, y) \parallel \min_{z \in C} f(x, y, z) \quad (36)$$

is jointly convex in x and y .

We proceed by induction on k . For the base case $k = 1$, $D^{(1)}(p \| q) = D(p \| q)$ is jointly strictly convex in $p, q \in \Delta$. Now assume the statement holds up to $k - 1$. Then $D(p \| w) + D^{(k-1)}(w \| q)$ is jointly strictly convex in $p, q, w \in \Delta$ and

$$D^{(k)}(p \| q) \parallel \min_{w \in \Delta} (D(p \| w) + D^{(k-1)}(w \| q)) \quad (37)$$

is jointly strictly convex in p, q by the above lemma (36) with $g(p, q) = D^{(k)}(p \| q)$ and $f(p, q, w) = D(p \| w) + D^{(k-1)}(w \| q)$.

Since $D^{(k)}(p \| q)$ is the minimum of a strictly convex function over a closed convex set Δ , there is a unique global minimum $(w_i^{(k)*})$ for $1 \leq i \leq k - 1$. $\square$

3) (Monotonicity of $D^{(k)}(p \| q)$ in k)

Consider the optimal k -step path $(q = w_0^{(k)*}, w_1^{(k)*}, \dots, w_{k-1}^{(k)*}, p = w_k^{(k)*})$ (15) and modify it by replacing $w_{k-1}^{(k)*}$ with p :

$$D^{(k)}(p \| q) = \sum_{i=1}^k D(w_i^{(k)*} \| w_{i-1}^{(k)*}) \quad (38)$$

$$\stackrel{(a)}{<} D(p \| p) + \sum_{i=1}^{k-1} D(w_i^{(k)*} \| w_{i-1}^{(k)*}) \quad (39)$$

$$\stackrel{(b)}{\leq} D^{(k-1)}(p||q) \quad (40)$$

(a) follows because we know from Theorem 2 that $w_{k-1}^{(k)*} \neq p$ and since we showed above that the optimal path $(w_i^{(k)*})$ is unique. (b) follows since in general $w_i^{(k-1)*} \neq w_i^{(k)*}$. $\square$

4) (Bounding $P_n^{(k)}(p|q)$)

First the upper bound (18). We proceed by induction on k . For $k = 1$, the statement holds (see [1] Theorem 11.1.4). Now assume the statement holds up to $k - 1$. Now using (9) we write

$$P_n^{(k)}(p|q) = \sum_{w \in \Delta_n} P_n(p|w) P_n^{(k-1)}(w|q) \quad (41)$$

$$\stackrel{(a)}{\leq} (n+1)^{(k-2)|\mathcal{X}|} \sum_{w \in \Delta_n} e^{-n(D(p||w) + D^{(k-1)}(w||q))} \quad (42)$$

$$\leq (n+1)^{(k-2)|\mathcal{X}|} |\Delta_n| \max_{w \in \Delta_n} e^{-n(D(p||w) + D^{(k-1)}(w||q))} \quad (43)$$

$$\stackrel{(b)}{\leq} (n+1)^{(k-1)|\mathcal{X}|} \max_{w \in \Delta} e^{-n(D(p||w) + D^{(k-1)}(w||q))} \quad (44)$$

$$= (n+1)^{(k-1)|\mathcal{X}|} e^{-n \min_{w \in \Delta} (D(p||w) + D^{(k-1)}(w||q))} \quad (45)$$

$$\stackrel{(c)}{=} (n+1)^{(k-1)|\mathcal{X}|} e^{-n D^{(k)}(p||q)} \quad (46)$$

where (a) follows from the inductive hypothesis, (b) follows from $|\Delta_n| \leq (n+1)^{|\mathcal{X}|}$ (see [1] Theorem 11.1.1) and from optimizing over Δ_n rather than Δ , and (c) follows from the recursive definition of the chained K-L divergence (11).

Next we find an asymptotic (in $n \rightarrow \infty$) lower bound. Let $(p_n)_{n \in \mathbb{N}}, p_n \in \mathcal{P}_n, \lim_{n \rightarrow \infty} p_n = p$.

Choose an n -indexed sequence $(v_{n,i})_{i=1}^{k-1}$ such that $v_{n,i} \in \mathcal{P}_n, D(v_{n,i}||w_i^*) \rightarrow 0$ as $n \rightarrow \infty$.

$n \longrightarrow \infty$. Such a sequence exists because $\bigcup_n \mathcal{X}_n$ is dense in $\mathcal{X}$. Denote $v_{n,k} \triangleq p_n$ and $w_k^* \| p$. Denote

$$\varepsilon_{n,i} \| v_{n,i} - w_i^* \quad (47)$$

By Pinsker's inequality

$$\lim_{n \rightarrow \infty} |\varepsilon_{n,i}[x]| = 0 \quad (48)$$

for all $i \in \{1, \dots, k\}$, $x \in \mathcal{X}$. Moreover, since $\lim_{n \rightarrow \infty} D(v_{n,i} \| w_i^*) = 0$ by assumption, then for large enough n we have $\text{Support}(v_{n,i}) = \text{Support}(w_i^*)$, so for large enough n

$$\text{Support}(\varepsilon_{n,i}) \subset \text{Support}(w_i^*) \quad (49)$$

Therefore for $i \in \{1, \dots, k\}$

$$\lim_{n \rightarrow \infty} D(v_{n,i} \| v_{n,i-1}) = \lim_{n \rightarrow \infty} \sum_{x \in \mathcal{X}} (w_i^*[x] + \varepsilon_{n,i}[x]) \log \left(\frac{w_i^*[x] + \varepsilon_{n,i}[x]}{w_{i-1}^*[x] + \varepsilon_{n,i-1}[x]} \right) \quad (50)$$

$$= \sum_{x \in \mathcal{X}} w_i^*[x] \log \left(\frac{w_i^*[x]}{w_{i-1}^*[x]} \right) \quad (51)$$

$$= D(w_i^* \| w_{i-1}^*) \quad (52)$$

where the second equality follows from (48) and (49) and $\varepsilon_n, 0[x] \triangleq 0$.

Using (9) we write

$$P_n^{(k)}(p_n | q) = \sum_{(w_i) \in \Delta_n^{k-1}} \prod_{i=1}^k P(w_i | w_{i-1}) \quad (53)$$

$$\stackrel{(a)}{\geq} \frac{1}{(n+1)^{k|\mathcal{X}|}} \sum_{(w_i) \in \Delta_n^{k-1}} e^{-n \sum_{i=1}^k D(w_i \| w_{i-1})} \quad (54)$$

$$\geq \frac{1}{(n+1)^{k|\mathcal{X}|}} e^{-n \min_{(w_i) \in \Delta_n^{k-1}} \sum_{i=1}^k D(w_i \| w_{i-1})} \quad (55)$$

with $w_0 = q$, $w_k = p_n$, $(w_i) = (w_1, \dots, w_{k-1})$, where (a) follows from the lower bound

$P_n(p|q) \geq e^{-nD(p||q)} / (n+1)^{|\mathcal{X}|}$ (see [1], Theorem 11.1.4). Therefore

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P_n^{(k)}(p_n|q) \geq \liminf_{n \rightarrow \infty} \left(-\frac{k|\mathcal{X}| \log(n+1)}{n} - \min_{(w_i) \in \Delta_n^{k-1}} \sum_{i=1}^k D(w_i \| w_{i-1}) \right) \quad (56)$$

$$= \liminf_{n \rightarrow \infty} \left(-\min_{(w_i) \in \Delta_n^{k-1}} \sum_{i=1}^k D(w_i \| w_{i-1}) \right) \quad (57)$$

$$\geq \liminf_{n \rightarrow \infty} \left(-\sum_{i=1}^k D(v_{n,i} \| v_{n,i-1}) \right) \quad (58)$$

$$\stackrel{(a)}{=} -\sum_{i=1}^k D(w_i^* \| w_{i-1}^*) \quad (59)$$

$$= -D^{(k)}(p||q) \quad (60)$$

where (a) follows by (52). Recalling the upper bound (18) we conclude

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P_n^{(k)}(p_n|q) = -D^{(k)}(p||q) \quad (61)$$

The first equality in (19) follows by letting $v_{n,k} = p_n = p$, $\epsilon_{n,k}[x] = 0$ for all n , $x \in \mathcal{X}$, so

$\lim_{n \rightarrow \infty} D(p||v_{n,k-1}) = D(p||w_{k-1}^*)$, so (59) holds, and

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P_n^{(k)}(p|q) = \lim_{n \rightarrow \infty} \frac{1}{n} \log P_n^{(k)}(p_n|q) = -D^{(k)}(p||q) \quad (62)$$

Equivalently, $P_n^{(k)}(p|q) \doteq e^{-nD^{(k)}(p||q)}$. $\square$

5) (Sanov's theorem)

This proof is very similar to the previous proof for the case $E = \{p\}$ and to the proof of Sanov's theorem in [1].

First the upper bound (20). Using (7) and (10) we write

$$P_n^{(k)}(E|q) = \sum_{p \in E \cap \Delta_n} \sum_{w \in \Delta_n} P_n(p|w) P_n^{(k-1)}(w|q) \quad (63)$$

$$\stackrel{(a)}{\leq} (n+1)^{(k-1)|\mathcal{X}|} \sum_{p \in E \cap \Delta_n} e^{-nD^{(k)}(p||q)} \quad (64)$$

$$\leq (n+1)^{(k-1)|\mathcal{X}|} \sum_{p \in E \cap \Delta_n} e^{-n \inf_{p \in E} D^{(k)}(p||q)} \quad (65)$$

$$= (n+1)^{(k-1)|\mathcal{X}|} \sum_{p \in E \cap \Delta_n} e^{-nD^{(k)}(E||q)} \quad (66)$$

$$\stackrel{(c)}{\leq} (n+1)^{k|\mathcal{X}|} e^{-nD^{(k)}(E||q)} \quad (67)$$

where (a) follows from the upper bound (18) and (b) follows from

$$|E \cap \Delta_n| \leq |\Delta_n| \leq (n+1)^{|\mathcal{X}|}.$$

By assumption E is closed and convex, so

$$p^{(k)*} = \underset{p \in E}{\operatorname{argmin}} D^{(k)}(p||q) \quad (68)$$

is the k -step I-projection of q on E (12). We find an asymptotic (in $n \rightarrow \infty$) lower bound.

Using (7) and (10) we write

$$P_n^{(k)}(E|q) = \sum_{p \in E \cap \Delta_n} \sum_{(w_i) \in \Delta_n^{k-1}} \prod_{i=1}^k P(w_i|w_{i-1}) \quad (69)$$

$$\stackrel{(a)}{\geq} \frac{1}{(n+1)^{k|\mathcal{X}|}} \sum_{p \in E \cap \Delta_n} \sum_{(w_i) \in \Delta_n^{k-1}} e^{-n \sum_{i=1}^k D(w_i \| w_{i-1})} \quad (70)$$

$$\geq \frac{1}{(n+1)^{k|\mathcal{X}|}} e^{-n \min_{p \in E \cap \Delta_n, (w_i) \in \Delta_n^{k-1}} \sum_{i=1}^k D(w_i \| w_{i-1})} \quad (71)$$

with $w_0 = q$, $w_k = p$, $(w_i) = (w_1, \dots, w_{k-1})$, where (a) follows from the lower bound

$$P_n(p|q) \geq e^{-nD(p||q)} / (n+1)^{|\mathcal{X}|} \text{ (see [1], Theorem 11.1.4).}$$

By assumption $E \cap \Delta_n$ is non-empty for sufficiently large n , so the lower bound (71) is not vacuous. We can then find a sequence $(p_n)_n$, $p_n \in E \cap \Delta_n$ and $\lim_{n \rightarrow \infty} D(p_n \| p^{(k)*}) = 0$ and an n -indexed sequence $(v_{n,i})_{i=1}^{k-1}$ such that $v_{n,i} \in \Delta_n$, $D(v_{n,i} \| w_i^*) \rightarrow 0$ as $n \rightarrow \infty$. Such a sequence exists because $\bigcup_n \Delta_n$ is dense in Δ . Recycling the argument in (52) we conclude

$$\lim_{n \rightarrow \infty} D(v_{n,i} \| v_{n,i-1}) = D(w_i^* \| w_{i-1}^*) \quad (72)$$

where $w_k^* \| p^{(k)*}$. Then

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P_n^{(k)}(E|q) \geq \liminf_{n \rightarrow \infty} \left(-\frac{k|\mathcal{X}| \log(n+1)}{n} - \min_{p \in E \cap \Delta_n} \min_{(w_i) \in \Delta_n^{k-1}} \sum_{i=1}^k D(w_i \| w_{i-1}) \right) \quad (73)$$

$$= \liminf_{n \rightarrow \infty} \left(-\min_{p \in E \cap \Delta_n} \min_{(w_i) \in \Delta_n^{k-1}} \sum_{i=1}^k D(w_i \| w_{i-1}) \right) \quad (74)$$

$$\geq \liminf_{n \rightarrow \infty} \left(-D(p_n \| v_{n,k-1}) - \sum_{i=1}^{k-1} D(v_{n,i} \| v_{n,i-1}) \right) \quad (75)$$

$$\stackrel{(a)}{=} -\sum_{i=1}^k D(w_i^* \| w_{i-1}^*) \quad (76)$$

$$= -D^{(k)}(E||q) \quad (77)$$

where (a) follows by (72).

Recalling the upper bound (20) we conclude

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P_n^{(k)}(E|q) = D^{(k)}(p^{(k)*}||q) \quad (78)$$

Equivalently, $P_n^{(k)}(E|q) \doteq e^{-nD^{(k)}(E||q)}$. $\square$

6) (Conditional limit theorem)

We follow closely the proof for the conditional limit theorem in [1]. Let

$(w_1^*, \dots, w_{k-1}^*) = \operatorname{argmin}_{(w_i) \in \Delta^{k-1}} D(p^{(k)*}||q)$ be the optimal path (15) and $p^{(k)*} = \operatorname{argmin}_{p \in E} D^{(k)}(p||q)$ (12) be the k -step I-projection of q on E ($p^{(k)*}$ exists and is unique since E is closed and convex by assumption).

Next define the set

$$S_t = \left\{ (w_1, \dots, w_{k-1}, p) \in \Delta^k : \sum_{i=1}^k D(w_i||w_{i-1}) \leq t \right\} \quad (79)$$

Let

$$D^* = \sum_{i=1}^k D_i^* = D^{(k)}(E||q) \quad (80)$$

Now define the set

$$A = S_{D^*+2\delta} \cap (\Delta^{k-1} \times E) \quad (81)$$

where the last term ensures that $(w_i)_{i=1}^k \in A \Rightarrow w_k \in E$. Define the set

$$B = (\Delta^{k-1} \times E) - A \quad (82)$$

Thus $A \cup B = \Delta^{k-1} \times E$. In the summations below,

$v = (w_1, \dots, w_{k-1}, w_k) \in \Delta_n^{k-1} \times (E \cap \Delta_n)$. Then

$$P_n^{(k)}(B|q) = \sum_{v: \sum_{i=1}^k D(w_i \| w_{i-1}) > D_* + 2\delta} \prod_{i=1}^k P(w_i | w_{i-1}) \quad (83)$$

$$\stackrel{(a)}{\leq} \sum_{v: \sum_{i=1}^k D(w_i \| w_{i-1}) > D_* + 2\delta} e^{-n \sum_{i=1}^k D(w_i \| w_{i-1})} \quad (84)$$

$$\leq \sum_{v: \sum_{i=1}^k D(w_i \| w_{i-1}) > D_* + 2\delta} e^{-n(D_* + 2\delta)} \quad (85)$$

$$\stackrel{(b)}{\leq} (n+1)^{k|\mathcal{X}|} e^{-n(D_* + 2\delta)} \quad (86)$$

where (a) follows from the upper bound (18) in the case $k = 1$ and (b) follows since

$$|\Delta_n^{k-1} \times (E_i \cap \Delta_n)| \leq |\Delta_n|^k \leq (n+1)^{k|\mathcal{X}|}.$$

In the summations below, $v = (w_1, \dots, w_{k-1}, w_k) \in \Delta_n^{k-1} \times (E \cap \Delta_n)$. We write

$$P_n^{(k)}(A|q) \stackrel{(a)}{\geq} P_n^{(k)}(A \cap S_{D_* + \delta} | q) \quad (87)$$

$$= \sum_{v: \sum_{i=1}^k D(w_i \| w_{i-1}) \leq D_* + \delta} \prod_{i=1}^k P(w_i | w_{i-1}) \quad (88)$$

$$\stackrel{(b)}{\geq} \sum_{v: \sum_{i=1}^k D(w_i \| w_{i-1}) \leq D_* + \delta} \frac{e^{-n \sum_{i=1}^k D(w_i \| w_{i-1})}}{(n+1)^{k|\mathcal{X}|}} \quad (89)$$

$$\stackrel{(c)}{\geq} \frac{e^{-n(D^*+\delta)}}{(n+1)^{k|\mathcal{X}|}} \text{ for sufficiently large } n \quad (90)$$

where in (a) we used $S_{D^*+\delta} \subset S_{D^*+2\delta}$ in (b) we used the lower bound

$P_n(p|q) \geq e^{-nD(p||q)} / (n+1)^{|\mathcal{X}|}$ (see [1], Theorem 11.1.4), and (c) follows since

$S_{D^*+\delta} \cap (\Delta^{k-1} \times E) \cap \Delta_n^k \neq \emptyset$ for n sufficiently large (since $E \cap \Delta_n^k \neq \emptyset$ for n sufficiently large by assumption).

Now recall the random variables $(X_t^{(i)})_{t \in \{1, \dots, n\}}$ defined in the statement of Theorem 1.6) and define the empirical distributions $\hat{w}_i \in \Delta_n^k$

$$\hat{w}_i[x] = \frac{1}{n} |\{t: x_t^{(i)} = x\}| \quad (91)$$

for all $x \in \mathcal{X}$, $i \in \{1, \dots, k\}$. Then for n sufficiently large

$$\mathbb{P}((\hat{w}_1, \dots, \hat{w}_k) \in B | \hat{w}_k \in E) = \frac{P_n^{(k)}(B \cap (\Delta^{k-1} \times E))}{P_n^{(k)}(\Delta^{k-1} \times E)} \quad (92)$$

$$\leq \frac{P_n^{(k)}(B)}{P_n^{(k)}(A)} \quad (93)$$

$$\stackrel{(a)}{\leq} \frac{(n+1)^{k|\mathcal{X}|} e^{-n(D^*+2\delta)}}{\frac{e^{-n(D^*+\delta)}}{(n+1)^{k|\mathcal{X}|}}} \quad (94)$$

$$= (n+1)^{2k|\mathcal{X}|} e^{-n\delta} \quad (95)$$

where (a) follows from (86) and (90). The above quantity goes to 0 as $n \rightarrow \infty$, so

$((\hat{w}_1, \dots, \hat{w}_k) \in A | \hat{w}_k \in E) \rightarrow 1$ as $n \rightarrow \infty$.

Next let's show that elements of A are close to $(w_1^*, \dots, w_k^*)$ in L^1 norm. This follows from the strict joint convexity of the function

$$\Lambda((w_i)_{i=1}^k) \left\| \sum_{i=1}^k D(w_i \| w_{i-1}) \right\| \quad (96)$$

on the compact, convex set $^{k-1} \times E$. We state a lemma:

Lemma 4—Let $f(x)$ be jointly strictly convex in $x \in C$, where C is a compact, convex subset of $\mathbb{R}^d$, let $x^* \triangleq \operatorname{argmin}_{x \in C} f(x)$. For $\delta > 0$ denote the sublevel set

$$T_\delta \triangleq \{x : f(x) - f(x^*) \leq \delta\} \quad (97)$$

Then

$$\lim_{\delta \rightarrow 0} \max_{x \in T_\delta} \|x - x^*\|_1 = 0 \quad (98)$$

Proof—Suppose (98) does not hold. Then there exists $\varepsilon > 0$ such that for all $\delta > 0$

$$\max_{x \in T_\delta} \|x - x^*\|_1 \geq \varepsilon \quad (99)$$

Let $T' \triangleq \{x : \|x - x^*\|_1 \geq \varepsilon\}$ and $\delta_\varepsilon \triangleq \min_{x \in T'} f(x) - f(x^*) > 0$ where the min exists by compactness of T' and the inequality follows since $x \in T' \Rightarrow x \neq x^*$ and $f(x)$ is jointly strictly convex. Then

$$x \in T_{\delta_\varepsilon/2} \Rightarrow f(x) - f(x^*) \leq \delta_\varepsilon/2 < \delta_\varepsilon \quad (100)$$

$$\Rightarrow x \in \overline{T'} \quad (101)$$

$$\Rightarrow \|x - x^*\|_1 < \varepsilon \quad (102)$$

This contradicts (99), so (98) holds. ■

The function $\Lambda((w_i)_{i=1}^k)$ is jointly strictly convex in $(w_i)_{i=1}^k$ with compact, convex support $^{k-1} \times E$ and attains its minimum at $\Lambda((w_i^*)_{i=1}^k) = D^{(k)}(E \| q)$, so by Lemma 4

$$\lim_{\delta \rightarrow 0} \max_{(w_i^*)_{i=1}^k \in A} \|(w_i)_{i=1}^* - (w_i^*)_{i=1}^k\|_1 = 0 \quad (103)$$

Now recalling our earlier conclusion that for all $\delta > 0$ $((\hat{w}_1, \dots, \hat{w}_k) \in A | \hat{w}_k \in E) \rightarrow 1$ as $n \rightarrow \infty$ and using (103), we conclude that for all $\varepsilon > 0$, all $i \in \{1, \dots, k\}$, and all $x \in \mathcal{X}$

$$\lim_{n \rightarrow \infty} (|\hat{w}_i[x] - w_i^*[x]| \geq \varepsilon | \hat{w}_k \in E) = 0 \quad (104)$$

□

B. Proof of Theorem 2

Proof

The proof of the local optimality condition (24) is given after the statement of this Theorem.

Now let's prove the geometric lower bound (25). We first state a lemma.

Lemma 5

Let $p, q \in \mathcal{P}$, $D(p||q) < \infty$, $w^*(p, q)$ be as in (23), and λ^* the value of the Lagrange multiplier so that $w^*(p, q)$ is normalized. Then

$$\lambda^* \leq 1 \quad (105)$$

Proof

First define the function

$$Z_\lambda(p, q) = \sum_{x \in \mathcal{X}} q[x] f_\lambda\left(\frac{p[x]}{q[x]}\right) \quad (106)$$

where

$$f_\lambda(t) = \begin{cases} \frac{-t}{W(te^\lambda)} + \frac{t}{W(e^\lambda)} & : t > 0 \\ -e^{-\lambda} & : t = 0 \end{cases} \quad (107)$$

We can check that $f_\lambda(\cdot)$ is convex and $f_\lambda(1) = 0$, so Z_λ defines an f -divergence. Now letting λ^* so that $w^*(p, q)$ (23) is normalized, we write

$$Z_{\lambda^*}(p, q) = \sum_{x \in \mathcal{X}} q[x] f_{\lambda^*} \left(\frac{p[x]}{q[x]} \right) \quad (108)$$

$$\stackrel{(a)}{=} \sum_{x \in \mathcal{X}} \left(-w * (p, q)[x] + \frac{p[x]}{W(e^{\lambda^*})} \right) \quad (109)$$

$$= -1 + \frac{1}{W(e^{\lambda^*})} \quad (110)$$

$$\stackrel{(b)}{\geq} 0 \\ \Downarrow \quad (111)$$

$$\lambda^* \leq 1 \quad (112)$$

where in (a) we substituted $w^*(p, q)$ (23) and in (b) we used the non-negativity of $\mathcal{F}$ divergences [ref?]. ■

Next let's state another lemma.

Lemma 6

Let $p, q \in \mathcal{P}$, $D(p||q) < \infty$ and $w^*(p, q)$ be as in (23). Then

$$w * (p, q)[x] \geq \sqrt{p[x]q[x]} \quad (113)$$

for all $x \in \mathcal{X}$.

Proof

Suppose there exists some $x \in \mathcal{X}$ such that (113) does not hold, so

$w * (p, q)[x] < \sqrt{p[x]q[x]}$. This implies $p[x], q[x] > 0$, so

$$w * (p, q)[x] = \frac{p[x]}{W(e^{\lambda^* \frac{p[x]}{q[x]}})} < \sqrt{p[x]q[x]} \\ \Downarrow \quad (114)$$

$$\lambda^* > \sqrt{\frac{p[x]}{q[x]}} - \log\left(\frac{p[x]}{q[x]}\right) \geq 1 \quad (115)$$

This contradicts Lemma 5, so the statement holds. ■

Lemma 6 implies

$$\log(w_i^{(k)*}[x]) = \log(w * (w_{i+1}^{(k)*}, w_{i-1}^{(k)*})[x]) \quad (116)$$

$$\geq \frac{1}{2} \log(w_{i+1}^{(k)*}[x]) + \frac{1}{2} \log(w_{i-1}^{(k)*}[x]) \quad (117)$$

Thus $\log w_i^{(k)*}[x]$ is concave in i , so

$$\log(w_i^{(k)*}[x]) \geq \frac{i}{k} \log(w_k^{(k)*}[x]) + \frac{k-i}{k} \log(w_0^{(k)*}[x]) \quad (118)$$

$$= \frac{i}{k} \log(p[x]) + \frac{k-i}{k} \log(q[x]) \quad (119)$$

so (25) follows. ■

C. Proof of Theorem 3

Proof

Let v_t denote the continuous path (31)

$$v_t \parallel \frac{1}{Z_{\tau_t}} (\tau_t \sqrt{p} + (1 - \tau_t) \sqrt{q})^2 \quad (120)$$

where Z_{τ_t} and τ_t are as in the statement of Theorem 3. Let $v_i^{(k)}$ denote the continuous path (31) sampled at k uniformly spaced points in time:

$$(v_i^{(k)})_{i=0}^k \parallel \left(\frac{1}{Z_{\tau_t}} (\tau_t \sqrt{p} + (1 - \tau_t) \sqrt{q})^2 \right)_{t \in \{0, \frac{1}{k}, \frac{2}{k}, \dots, 1\}} \quad (121)$$

Let $(w_i^{(k)*})_{i=0}^k$ denote the optimal k -step path (15). Then $v_0^{(k)} = w_0^{(k)*} = q$ and $v_k^{(k)} = w_k^{(k)*} = p$. For $1 \leq i \leq k$ denote by $\theta_i^{(k)}$ the angle between (the square roots of) adjacent distributions on the path $(v_i^{(k)})_{i=0}^k$

$$\phi_i^{(k)} \parallel \arccos \left(\sqrt{v_i^{(k)}} \cdot \sqrt{v_{i-1}^{(k)}} \right) = \frac{1}{k} \theta_{\sqrt{p}, \sqrt{q}} \quad (122)$$

$$\theta_i^{(k)} \parallel \arccos \left(\sqrt{w_i^{(k)*}} \cdot \sqrt{w_{i-1}^{(k)*}} \right) \quad (123)$$

Next let

$$\mu \parallel \min_{x \in \text{Support}(q)} \min(p[x], q[x]) > 0 \quad (124)$$

We have $\mu > 0$ since $D(p \parallel q) < \infty$ and $D(q \parallel p) < \infty$, so $\text{Support}(p) = \text{Support}(q)$ by assumption.

Now apply the series expansion for the K-L divergence (27) to $(v_i^{(k)})_{i=0}^k$ and $(w_i^{(k)*})_{i=0}^k$ and sum over $i \in \{1, \dots, k\}$:

$$\sum_{i=1}^k D(v_i^{(k)} \parallel v_{i-1}^{(k)}) = \sum_{i=1}^k 2 \left(\phi_i^{(k)2} + O(\phi_i^{(k)4}) + \sum_{x \in \mathcal{X}} O \left(\frac{\phi_i^{(k)3}}{\sqrt{v_i^{(k)}}} \right) \right) \quad (125)$$

$$= \sum_{i=1}^k \left(2\phi_i^{(k)2} + O(\phi_i^{(k)3}) \right) \quad (126)$$

where the second equality comes from the i - and k - independent lower bound

$$\min_{x \in \mathcal{X}} \min_{t \in [0,1]} v_t[x] \geq \mu \parallel \min_{x \in \mathcal{X}} \min(p[x], q[x]) > 0 \quad (127)$$

Since $D(p \parallel q) < \infty$ and $D(q \parallel p) < \infty$ by assumption, we may assume $p[x], q[x] > 0$ for all $x \in \mathcal{X}$ (since otherwise we could restrict attention to $\text{Support}(p) = \text{Support}(q) \subset \mathcal{X}$). Thus the last inequality in (127) follows. Similarly we obtain

$$\sum_{i=1}^k D(w_i^{(k)*} \| w_{i-1}^{(k)*}) = \sum_{i=1}^k 2 \left(\theta_i^{(k)2} + O(\theta_i^{(k)4}) + \sum_{x \in \mathcal{X}} O \left(\frac{\theta_i^{(k)3}}{\sqrt{w_i^{(k)*}}} \right) \right) \quad (128)$$

$$= \sum_{i=1}^k \left(2\theta_i^{(k)2} + O(\theta_i^{(k)3}) \right) \quad (129)$$

where the last inequality follows from the i - and k -independent lower bound

$$\min_{x \in \mathcal{X}} \min_{i \in \{0, \dots, k\}} w_i^{(k)*} \geq \min_{x \in \mathcal{X}} p[x]^{i/k} q[x]^{(k-i)/k} \geq \mu > 0 \quad (130)$$

where the first inequality follows from Theorem 2, (25).

Now we write

$$\lim_{k \rightarrow \infty} k \sum_{i=1}^k \left(D(w_i^{(k)*} \| w_{i-1}^{(k)*}) - D(v_i^{(k)} \| v_{i-1}^{(k)}) \right) = \lim_{k \rightarrow \infty} 2k \sum_{i=1}^k \left(\theta_i^{(k)2} - \phi_i^{(k)2} \right) \quad (131)$$

$$= \lim_{k \rightarrow \infty} 2k \sum_{i=1}^k \left(\theta_i^{(k)2} - \frac{\theta^2}{k^2} \right) \quad (132)$$

$$\geq 0 \quad (133)$$

where the inequality follows since the path v_t is part of a great circle joining $\sqrt{p}$ and $\sqrt{q}$ at constant angular speed (uniformly spacing the angles $\phi_i^{(k)} = \frac{1}{k} \theta_{\sqrt{p}, \sqrt{q}}$ minimizes the sum of their squares subject to their sum $\phi_i^{(k)} = \theta_{\sqrt{p}, \sqrt{q}}$). On the other hand by the optimality of $(w_i^*)_{i=1}^k$

$$\lim_{k \rightarrow \infty} k \sum_{i=1}^k \left(D(w_i^{(k)*} \| w_{i-1}^{(k)*}) - D(v_i^{(k)} \| v_{i-1}^{(k)}) \right) \leq 0 \quad (134)$$

Thus, combining the above inequalities,

$$\lim_{k \rightarrow \infty} k \sum_{i=1}^k \left(D(w_i^{(k)*} \| w_{i-1}^{(k)*}) - D(v_i^{(k)} \| v_{i-1}^{(k)}) \right) = 0 \quad (135)$$

We then have

$$\sum_{i=1}^k \left(D(w_i^{(k)*} \| w_{i-1}^{(k)*}) - D(v_i^{(k)} \| v_{i-1}^{(k)}) \right) = \sum_{i=1}^k (O(\theta_i^{(k)3}) + O(\phi_i^3)) \quad (136)$$

$$= k O(1/k^3) = O(1/k^2) \quad (137)$$

[why $\theta^{(k)} = O(1/k)$]

[Now use joint strict convexity of $\sum_{i=1}^k D(w_i \| w_{i-1})$ on the compact, convex set $^{k-1} \times E$ to complete proof?] ■

References

1. Cover, TM., Thomas, JA. Elements of Information Theory. Second. Hoboken, NJ: John Wiley & Sons; 2006.
2. Wright S. Evolution in Mendelian populations. Genetics. 1931; 16:97–159. [PubMed: 17246615]
3. Crow JF. Wright and Fisher on inbreeding and random drift. Genetics. 2010; 184:609–611. [PubMed: 20332416]
4. Sanov IN. On the probability of large deviations of a random variable. Mat Sbornik. 1957; 42:11–44.
5. Csiszár I. Sanov property, generalized I-projection and a conditional limit theorem. The Annals of Probability. 1984; 12:768–793.
6. Csiszár I, František M. Information projections revisited. IEEE Transactions on Information Theory. 2003; 49:1474–1490.
7. Papangelou F. The large deviations of a multi-allele Wright-Fisher process mapped on the sphere. The Annals of Applied Probability. 2000; 10:1259–1273.
8. Pavlichin, DS., Weissman, T. Chained Kullback-Leibler divergences. www.stanford.edu/~dmitrip/chaineddivergences.pdf, in preparation
9. Boyd, S., Vandenberghe, L. Convex Optimization. Cambridge, UK: Cambridge University Press; 2009.
10. Hájek J. Asymptotically most powerful rank-order tests. The Annals of Mathematical Statistics. 1962; 33:1124–1147.
11. Le Cam L. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. The Annals of Mathematical Statistics. 1970; 41:802–828.
12. Pollard, D. Another look at differentiability in quadratic mean. In: Pollard, GLYD., Torgersen, E., editors. Festschrift for Lucien Le Cam. New York: Springer; 1997. p. 305–314. ch 9
13. Amari, S., Nagaoka, H. Methods of information geometry. American Mathematical Society; 2000.

Fig. 1.

(left) The Wright-Fisher process, a Markov chain on the set of types. (right) A trajectory of the Wright-Fisher process for $|\mathcal{X}|=3$ alleles plotted on the simplex. Eventually only one allele remains (“fixes”).

Fig. 2.

(left) k -step path (15) from q to p (corresponding to $D^{(k)}(p||q)$) computed by successively applying (24) for $w_1^{(k=2)*}$ and $\left(w_i^{(k=6)*}\right)_{1 \leq i \leq 5}$. The green line shows the limiting $k \rightarrow \infty$ path (31). The dark grey lines show the arithmetic and (normalized) geometric mean of p and q for $\alpha \in [0, 1]$. (right) The limiting path as $k \rightarrow \infty$, part of a great circle joining $\sqrt{p}$ and $\sqrt{q}$.

Fig. 3.

(left) Maxwell's demon stands ready to release particles into the water supply below if the distribution of a random sample of n happens to be in E . (right) Two demons use an intermediate stage E' to desalinate faster.