Appendix S1: Proofs of properties of the DSD and maximizers

Jonathan M. Henshaw & Yoav Zemel

This is supporting information for the article:

> Henshaw, J.M. & Zemel, Y. (2016). A unified measure of linear and nonlinear selection on quantitative traits. *Methods in Ecology and Evolution.* doi: 10.1111/2041-210X.12685.

It contains three sections. In Section 1 we show how the maximizers h are constructed, and derive eqn 17 in the main text. In Section 2 we decompose the distributional selection differential (DSD) d into a directional component and a nondirectional component. In Section 3, we show that the overall DSD across multiple episodes of selection is less than or equal to the sum of the DSDs for each episode.

Trait values before selection are represented by a random variable Z with probability distribution $\mathcal{P}$ and cumulative distribution function G. Trait values after selection are represented similarly by Z^*, $\mathcal{P}^*$ and G^*. If h is any function of trait values, its expectations before and after selection are written $\mathbb{E}h(Z)$ and $\mathbb{E}h(Z^*)$, i.e.

$$\mathbb{E}h(Z) = \int_{-\infty}^{\infty} h(z) d\mathcal{P}(z), \qquad \mathbb{E}h(Z^*) = \int_{-\infty}^{\infty} h(z) d\mathcal{P}^*(z).$$

For example, if trait values are continuous random variables with densities g and g^*, then we have

$$\mathbb{E}h(Z) = \int_{-\infty}^{\infty} h(z) g(z) dz, \qquad \mathbb{E}h(Z^*) = \int_{-\infty}^{\infty} h(z) g^*(z) dz.$$

Similarly, if Z and Z^* are discrete random variables concentrated on a finite set of points $z_1, \ldots, z_n$ with weights $p_i = p(z_i)$ and $p_i^* = p^*(z_i)$, then

$$\mathbb{E}h(Z) = \sum_{i=1}^{n} h(z_i) p_i, \qquad \mathbb{E}h(Z^*) = \sum_{i=1}^{n} h(z_i) p_i^*.$$

Most of this supplement applies to any two arbitrary trait distributions Z and Z^*. In the main text, we follow the usual convention that the trait distribution after selection is given by $d\mathcal{P}^* = \mathbb{E}(w|Z)\, d\mathcal{P}$, where w is relative fitness. In this case, $g^*(z) = \mathbb{E}(w|z) g(z)$ and $p^*(z) = \mathbb{E}(w|z) p(z)$.

1 Explicit construction of maximizers h

In this section we identify the class of maximizers h that arises in the covariance definition of the DSD (eqns 15–16 in main text) and which are used to define distributional selection gradients (eqn 27 in main text). First, we recall some definitions and terminology from the main text. Next, we construct h for the intuitive case of discrete distributions. Last, we provide the fully general construction for the interested reader.

1.1 Definitions

A function h satisfies the *gradient condition* (denoted $h \in \text{Grad}$) if for any two values z_1 and z_2 (eqn 14 in main text),

$$|h(z_1) - h(z_2)| \leq |z_1 - z_2|.$$

In mathematics, such functions are called Lipschitz (or, more precisely, 1-Lipschitz) functions (Villani 2009). When h is differentiable, $h \in \text{Grad}$ if and only if $|h'(z)| \leq 1$ for all z, hence the term 'gradient condition'.

A *maximizer* is a function h that maximizes the quantity $\mathbb{E}h(Z^*) - \mathbb{E}h(Z)$ while satisfying the gradient condition. From the cumulative integral definition of the DSD (eqn 18 in main text), we know that the maximal value is given by

$$d = \max_{h \in \text{Grad}} (\mathbb{E}h(Z^*) - \mathbb{E}h(Z)) = \int_{-\infty}^{\infty} |G(z) - G^*(z)| \, dz. \tag{S1}$$

Any maximizer h must obtain the value on the right-hand side exactly.

1.2 Construction of maximizers in the discrete case

The construction of maximizers is more transparent in the following setup: suppose that trait values before and after selection are concentrated on a finite number of points $z_1 \leq \cdots \leq z_n$ on the real line, with weights p_i and p_i^* respectively. In this case, (S1) simplifies to

$$d = \sum_{i=1}^{n-1} (z_{i+1} - z_i) \left| \sum_{j=1}^{i} p_j - p_j^* \right|. \tag{S2}$$

This matches eqn 19 in the main text when $p_j = 1/n$ and $p_j^* = p_j w_j$. In this setting, the expression to be maximized is

$$\begin{aligned}\mathbb{E}h(Z^*) - \mathbb{E}h(Z) &= \sum_{j=1}^{n} h(z_j)(p_j^* - p_j) \\ &= \sum_{i=1}^{n-1} [h(z_{i+1}) - h(z_i)] \sum_{j=1}^{i} p_j - p_j^*.\end{aligned} \tag{S3}$$

For each i, an optimal choice for $h(z_{i+1}) - h(z_i)$ given the gradient condition $|h(z_{i+1}) - h(z_i)| \leq z_{i+1} - z_i$ is

$$h(z_{i+1}) - h(z_i) = (z_{i+1} - z_i) \operatorname{sgn}\left(\sum_{j=1}^{i} p_j - p_j^*\right), \tag{S4}$$

where sgn is the sign function

$$\operatorname{sgn}(x) = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \\ 0, & x = 0 \end{cases}.$$

Once values for the differences $h(z_{i+1}) - h(z_i)$ have been chosen for each $i = 1, \ldots, n-1$, the values of all $h(z_i)$ are determined uniquely by the choice of $h(z_1)$. This choice is arbitrary, as $h(z_1)$ does not appear in the objective function (S3).

We then interpolate h to be linear on each interval $[z_i, z_{i+1}]$ to ensure that the gradient condition is met. (If it is desirable to have a function h defined throughout the real line, one can also set $h(z) = h(z_1)$ for $z < z_1$ and $h(z) = h(z_n)$ for $z > z_n$.) For this choice of h, the right-hand side of (S3) matches (S2), confirming optimality. Furthermore, the calculation above shows that whenever $\sum_{j=1}^{i} p_j^* - p_j \neq 0$, the value of $h(z_{i+1}) - h(z_i)$ from (S4) is uniquely optimal. If $\sum_{j=1}^{i} p_j^* - p_j = 0$ for some i, then $h(z_{i+1}) - h(z_i)$ can take any value between $-(z_{i+1} - z_i)$ and $(z_{i+1} - z_i)$, and the optimal h is not unique. In either case, shifting h by a constant does not affect the value of $\mathbb{E}h(Z^*) - \mathbb{E}h(Z)$, so in the best case scenario h is unique up to an additive constant. In the main text, we suggest standardising h by taking $\operatorname{sgn}(0) = 0$ and by choosing $h(z_1)$ so that $\mathbb{E}h(Z) = 0$.

1.3 General case

We now turn to the general case where trait values before and after selection are represented by any distribution functions G and G^* on $\mathbb{R}$. In this case, a maximizer is any function h of the form (eqn 17 in main text):

$$h(z) = h(0) + \int_0^z \operatorname{sgn}(G(x) - G^*(x))\, dx.$$

(In the discrete setup, G and G^* are constant on each interval $[z_i, z_{i+1})$, and we recover the result of the previous subsection.)

The proof below is similar to the discrete case, with sums replaced by integrals and differences replaced by derivates. We assume that trait values before and after selection have well-defined mean values (i.e. $\mathbb{E}|Z| < \infty$ and $\mathbb{E}|Z^*| < \infty$). This is a technical condition that will be met in any empirical application, and without which the DSD may be infinite.

Any h satisfying the gradient condition can be written as the integral of its derivative (since Lipschitz functions are absolutely continuous: Stein & Shakarchi 2005). Without loss of generality, we can take $h(0) = 0$, in which case

$$h(z) = \int_0^z h'(t)\, dt. \tag{S5}$$

We follow the usual convention that if $z < 0$ then $\int_0^z h'(t)\, dt = -\int_z^0 h'(t)\, dt$. The absolute value of the derivate is bounded by one (i.e. $|h'(t)| \leq 1$ for all t), so that $|h(Z)| \leq |Z|$ has finite expectation. Consequently, all of the integrals in the following calculation are finite (in absolute value) and Fubini's theorem (Stein & Shakarchi 2005) implies that

$$\begin{aligned}
\mathbb{E}h(Z) &= \int_{-\infty}^{\infty} h(z)\, d\mathcal{P}(z) \\
&= \int_0^{\infty} d\mathcal{P}(z) \int_0^z h'(t)\, dt - \int_{-\infty}^0 d\mathcal{P}(z) \int_z^0 h'(t)\, dt \\
&= \int_0^{\infty} h'(t)\, dt \int_t^{\infty} d\mathcal{P}(z) - \int_{-\infty}^0 h'(t)\, dt \int_{-\infty}^t d\mathcal{P}(z) \\
&= \int_0^{\infty} h'(t)(1 - G(t))\, dt - \int_{-\infty}^0 h'(t)G(t)\, dt.
\end{aligned} \tag{S6}$$

Repetition of this calculation for $\mathbb{E}h(Z^*)$ and then subtraction yields:

$$\mathbb{E}h(Z^*) - \mathbb{E}h(Z) = \int_{-\infty}^{\infty} h'(t)\left(G(t) - G^*(t)\right)\, dt. \tag{S7}$$

Given that $|h'(t)| \leq 1$, the right-hand side of this equation is maximized when $h'(t) = \operatorname{sgn}(G(t) - G^*(t))$ for all t. The value of $\mathbb{E}h(Z^*) - \mathbb{E}h(Z)$ thus obtained is the same as in (S1), confirming the optimality of this solution. Uniqueness holds provided that the set $E = \{t : G(t) = G^*(t)\}$ of points of ambiguity is small enough. For example, if traits are continuously distributed and the graphs of G and G^* intersect only finitely many times, then

h is unique up to an additive constant. More generally, h is unique up to an additive constant if E has total length (i.e. Lebesgue measure) zero. As above, we propose setting $h' = 0$ on E and then standardising h so that $\mathbb{E}h(Z) = 0$.

2 Decomposition of the DSD into directional and nondirectional components

Here we show that any flow can be decomposed into two components: a directional flow that shifts trait values entirely in one direction, and a nondirectional flow that reshapes the trait distribution without changing the trait mean. This allows the DSD to be partitioned into a directional component $d_D = |s|$ and a nondirectional component $d_N = d - |s|$.

First we fix some terminology. Let F be a flow between trait distributions represented by random variables Z and Z^* (i.e. F is a joint probability distribution of Z and Z^*). We write $d_F = \mathbb{E}_F|Z^* - Z|$ for the total amount of work associated with F and $s_F = \mathbb{E}(Z^* - Z)$ for the change in mean trait values. (The latter does not actually depend on the flow F, but the notation will help to remind us which distributions are involved.) We call a flow *directional* if $d_F = |s_F|$, meaning that it moves mass entirely in one direction (i.e. from lower to higher trait values or vice versa). A *nondirectional* flow leaves the trait mean unchanged, so that $s_F = 0$.

Suppose Z and Z^* are the trait distributions before and after selection and F is an optimal flow between them. We write $d = d_F$ for the associated DSD and $s = s_F$ for the linear selection differential. We will construct an 'intermediate' trait distribution Q and a joint distribution of (Z, Q, Z^*) such that

(i) the marginal distribution of (Z, Z^*) equals the original flow F;

(ii) the marginal distribution of (Z, Q) is a directional flow D, which satisfies $s_D = s$ and $d_D = |s|$; and

(iii) the marginal distribution of (Q, Z^*) is a nondirectional flow N, which satisfies $s_N = 0$ and $d_N = d - |s|$.

Since $d = d_D + d_N$, we can think of this as a decomposition of F into two sequential flows, corresponding to two consecutive episodes of selection. The first episode shifts the trait distribution from Z to Q via the directional flow D. The second episode shifts the trait distribution from Q to Z^* via the nondirectional flow N. The component flows D and N are also optimal, as otherwise the optimality of F would be contradicted.

2.1 Uphill and downhill flows

The original flow F can be split into an 'uphill' flow that moves mass from smaller to larger trait values, and a 'downhill' flow that moves mass in the other direction. The supports of these flows are $H = \{Z < Z^*\}$ and $L = \{Z > Z^*\}$ respectively. We refer to mass left in place by F as the 'diagonal' flow.

The total work associated with the uphill and downhill components of F is $h = \mathbb{E}_F(Z^* - Z)\chi_H$ and $\ell = \mathbb{E}_F(Z - Z^*)\chi_L$, where χ_A is the indicator function of a set A. The diagonal flow of course requires no work at all. The DSD equals the sum of the uphill and downhill work (i.e. $d = h + \ell$), whereas the linear selection differential is their difference (i.e. $s = h - \ell$).

2.2 Constructing the component flows D and N

We now construct flows D and N with the desired properties. Without loss of generality, we will assume that $s \geq 0$, meaning that uphill flow is at least as costly as downhill flow under F. The proof when $s < 0$ is very similar.

The original flow F defines a joint distribution of (Z, Z^*). We define the random variable Q as the mixture of $\max(Z, Z^*)$ and Z with probabilities s/h and $1 - s/h$ respectively. This defines a joint distribution of (Z, Q, Z^*). We can then take the marginal distribution of (Z, Q) as the directional flow D and the marginal distribution of (Q, Z^*) as the nondirectional flow N. Intuitively, D carries a proportion s/h of the original uphill flow but otherwise leaves all mass in place, and N completes the original flow by carrying the remaining uphill flow and all of the downhill flow.

2.3 DSDs and linear selection differentials for D and N

The DSDs associated with the two component flows are given by

$$\begin{aligned}
d_D &= \mathbb{E}_D|Q - Z| \\
&= \left(\frac{s}{h}\right) \mathbb{E}_F|\max(Z, Z^*) - Z| + \left(1 - \frac{s}{h}\right) \mathbb{E}_F|Z - Z| \\
&= \left(\frac{s}{h}\right) \mathbb{E}_F(Z^* - Z)\chi_H \qquad \text{(S8)} \\
&= \left(\frac{s}{h}\right) h \\
&= s
\end{aligned}$$

and

$$\begin{aligned} d_N &= \mathbb{E}_N|Z^* - Q| \\ &= \left(\frac{s}{h}\right)\mathbb{E}_F|Z^* - \max(Z, Z^*)| + \left(1 - \frac{s}{h}\right)\mathbb{E}_F|Z^* - Z| \\ &= \left(1 - \frac{s}{h}\right)\mathbb{E}_F(Z^* - Z)\chi_H + \mathbb{E}_F(Z - Z^*)\chi_L \\ &= \left(1 - \frac{s}{h}\right)h + \ell \\ &= d - s. \end{aligned} \tag{S9}$$

(If we had started by assuming that $s < 0$ then we would have $d_D = -s$ and $d_N = d + s$.) By similar arguments, the linear selection differentials are (always) given by $s_D = s$ and $s_N = 0$.

3 DSD across multiple episodes of selection

It is sometimes informative to break down fitness into several multiplicative components (e.g. reproductive lifespan and average rate of reproduction), so-called *episodes of selection* (Arnold & Wade 1984). Suppose there are m episodes, with relative success in the ith episode written as w_i. Relative fitness across all episodes is then $w = \prod_{i=1}^m w_i$. Writing s_i for the linear selection differential associated with the ith episode, the total selection differential is simply $s = \sum_{i=1}^m s_i$ (Arnold & Wade 1984). In contrast, the DSD measures the absolute change in trait distributions for each episode of selection. If selection acts in opposing directions among episodes, these changes may 'cancel out' to give a total change less than the sum of the component changes.

Let us write Z_0 for the (random variable representing) trait distribution before selection and Z_i for the trait distribution after the ith episode of selection. Let d_i be the DSD between Z_{i-1} and Z_i, with associated optimal flow F_i. Similarly, let d be the overall DSD between Z_0 and Z_m, with optimal flow F. We will show that

$$d \leq \sum_{i=1}^{m} d_i. \tag{S10}$$

The proof follows from the earth mover's definition of the DSD (eqn 13 in main text). Each of the optimal flows F_i is a joint distribution of (Z_{i-1}, Z_i). We can 'glue together' these distributions into a joint distribution F^* of $(Z_0, Z_1, \ldots, Z_m)$, such that the marginal distributions of F^* match the F_i in the obvious way (see the 'Gluing Lemma' in Villani 2009). The marginal distribution of (Z_0, Z_m) in F^* is a flow between Z_0 and Z_n, but it need not match F. Since F is an optimal flow, however, we are guaranteed that

$$d = \mathbb{E}_F |Z_m - Z_0| \leq \mathbb{E}_{F^*} |Z_m - Z_0|. \tag{S11}$$

Further, by the triangle inequality we have

$$\mathbb{E}_{F^*} |Z_m - Z_0| \leq \sum_{i=1}^{m} \mathbb{E}_{F^*} |Z_i - Z_{i-1}| = \sum_{i=1}^{m} d_i. \tag{S12}$$

Combining (S11) and (S12) gives the required result.

References

Arnold, S.J. & Wade, M.J. (1984). *On the measurement of natural and sexual selection: theory.* Evolution, **38**, 709–719.

Stein, E.M. & Shakarchi, R. (2005) *Real Analysis: Measure Theory, Integration & Hilbert Spaces.* Princeton University Press, Princeton.

Villani, C. (2009). *Optimal Transport: Old and New.* Springer, Berlin.