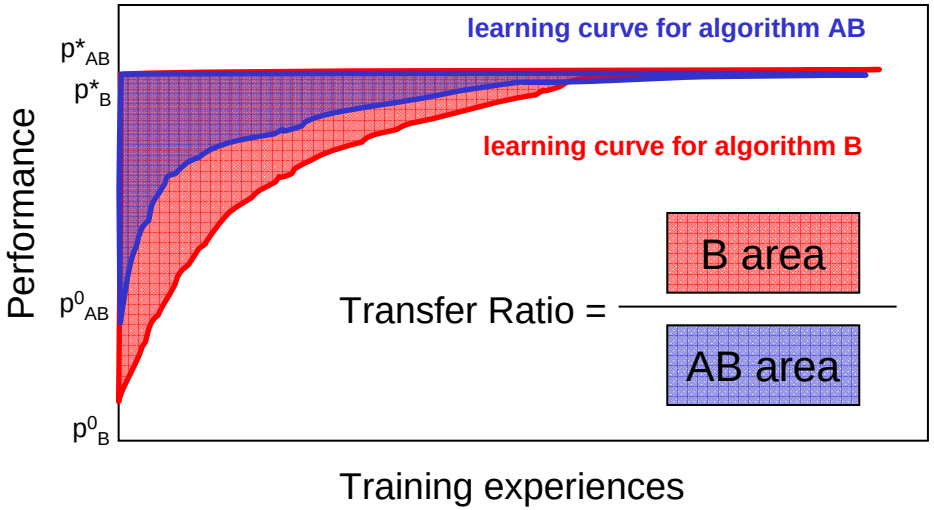
Proposed Metrics for Transfer Learning

Version 4
April 30, 2006

Tom Dietterich, editor

Shortcomings of Transfer Ratio



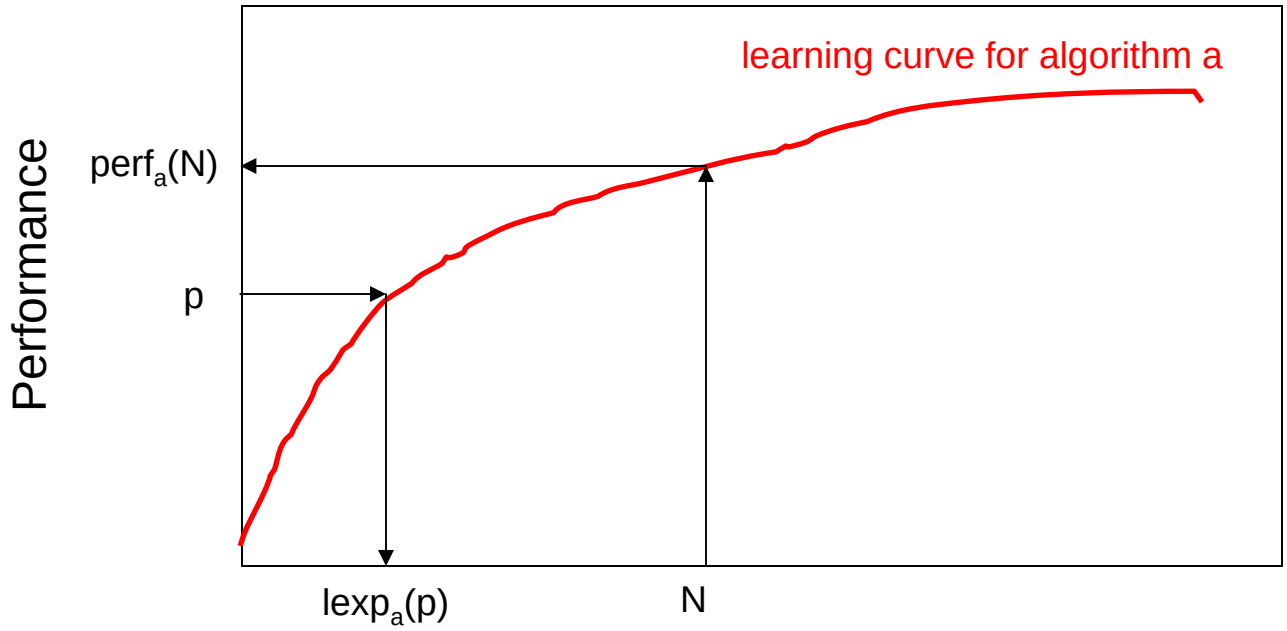
- Hard to explain: no well-defined units
- Large dependence on correctly estimating the asymptotic performance p^*_{AB}
- Infinite when $p^*_{AB} > p^*_{B}$
- Does not give an instantaneous value at each sample size. This can hide many interesting phenomena
- Alternate minimum TR metric only measures the worst-case. Can hide excellent performance at small-to-medium sample sizes

Proposed Metrics

- Average speedup
  - advantage: easy to interpret
  - disadvantage: can be infinite even after removing obvious cases
- Average relative reduction in amount of training data required
  - advantage: also easy to interpret
  - advantage: avoids most infinite answers

Notation

- Let p = $perf_{alg}(N)$ be the performance of algorithm "alg" after N training "experiences"
- Let N = $lexp_{alg}(p)$ be the inverse function: the number of training experiences N required to achieve target level p



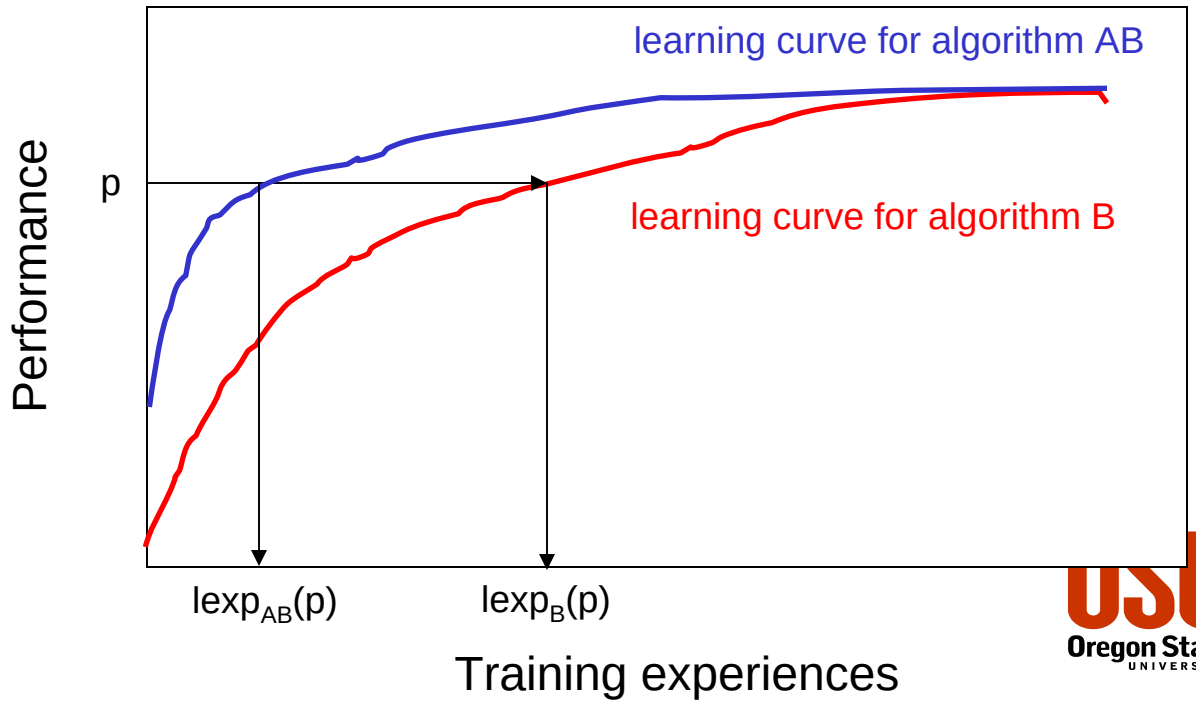
Learning Speedup and Relative Reduction

- Let
  - algorithm B = learning without transfer
  - algorithm AB = learning with transfer
- The speedup in the number of training examples required to reach target performance level P is

$$\text{speedup}(p) = \frac{\text{lexp}_B(p)}{\text{lexp}_{AB}(p)}$$

- The relative reduction in amount of training required is

$$\text{RR}(p) = \frac{\text{lexp}_B(p) - \text{lexp}_{AB}(p)}{\text{lexp}_B(p)} = 1 - \frac{\text{lexp}_{AB}(p)}{\text{lexp}_B(p)}$$



Statistical Interpretation of Speedup

- The asymptotic value of the ratio

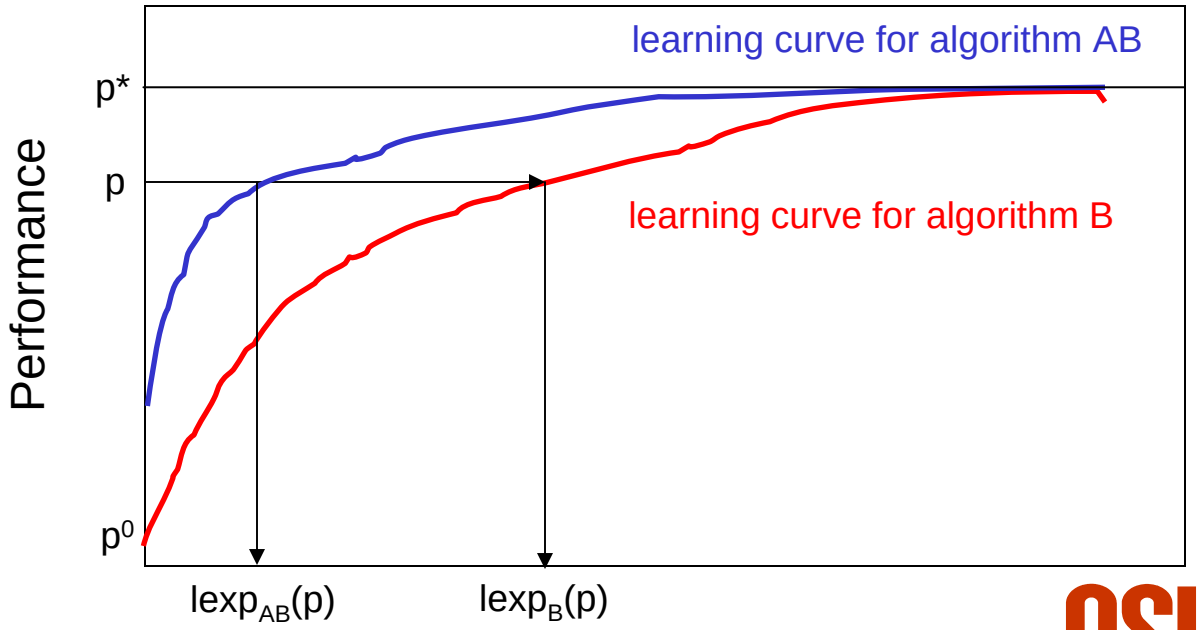
$$\lim_{p \to p^*} \frac{\mathrm{lexp}_B(p)}{\mathrm{lexp}_{AB}(p)}$$

is known as the statistical efficiency of algorithm AB relative to algorithm A. It is a standard measure employed in statistics to compare *consistent* learning algorithms – that is, algorithms that reach optimal performance asymptotically

Integrated Speedup

- Goal: obtain a single number that summarizes the relative performance.
- Solution: Integrate this speedup p^0 up to the asymptote for the task:

$$\frac{1}{p^* - p^0} \int_{p^0}^{p^*} \text{speedup}(p)\,dp = \frac{1}{p^* - p^0} \int_{p^0}^{p^*} \frac{\text{lexp}_B(p)}{\text{lexp}_{AB}(p)} dp$$



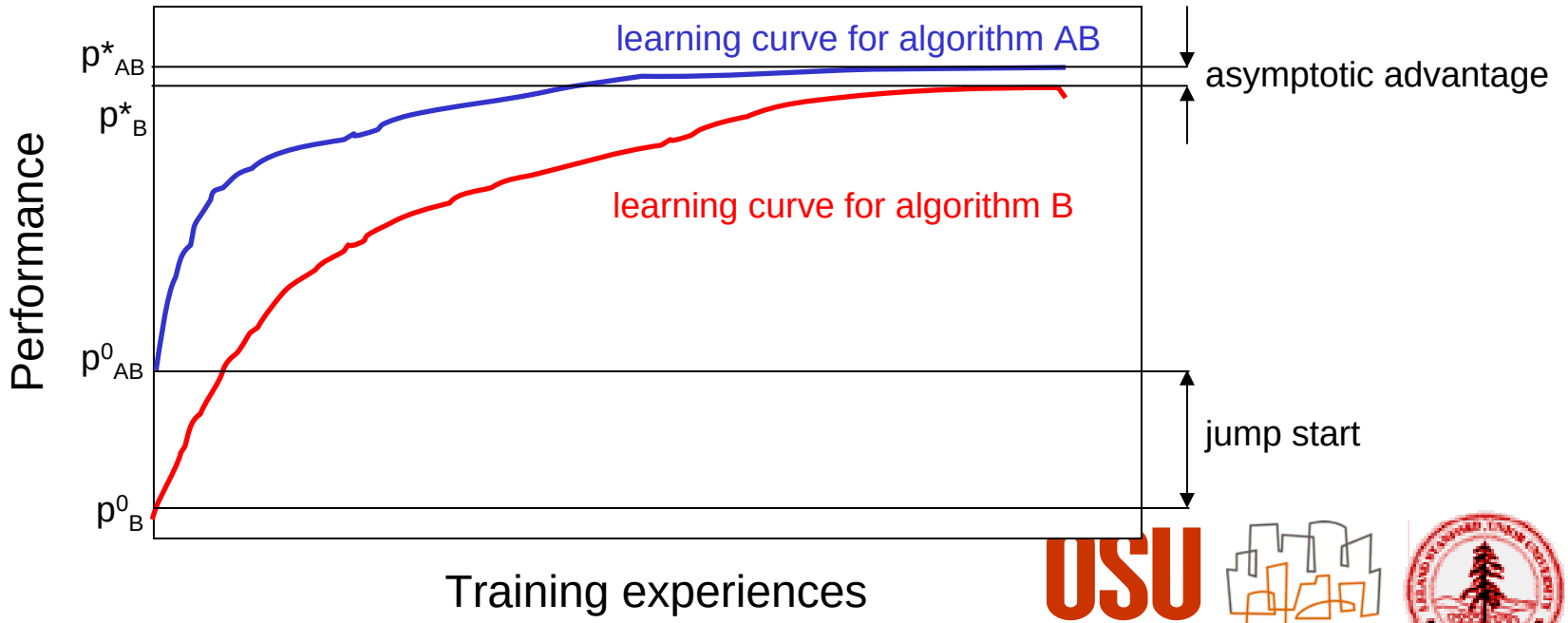
Issues

- The ratio will be infinite if the AB curve has a Y-axis intercept above the B curve intercept
  - at such points, we have infinite speedup
- The ratio will be infinite if the AB curve has an asymptote above the B curve asymptote
  - for levels p > the B curve asymptote, we have infinite speedup

Partial Solution: Remove the Problem Regions

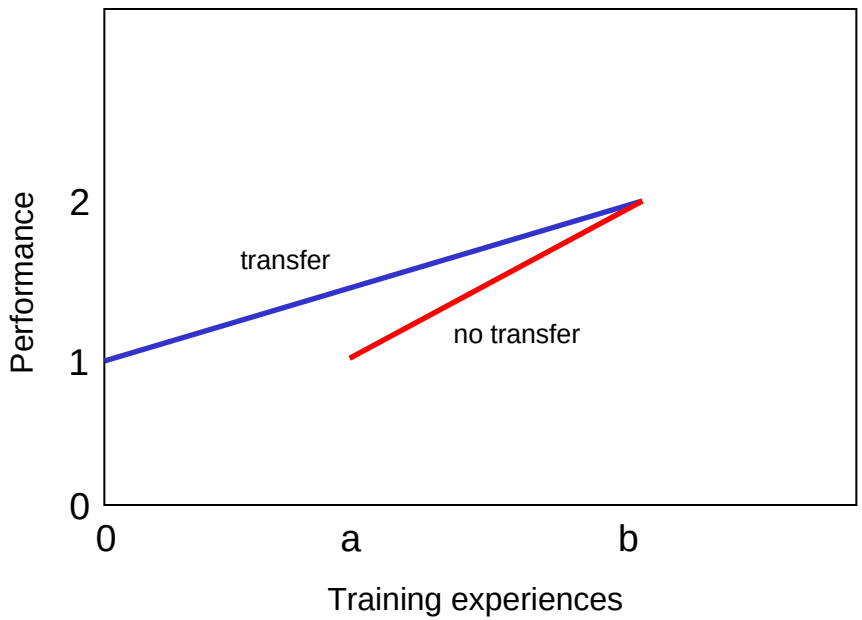
- Define three quantities:
  - "jump start": $p^0_{AB} - p^0_B$
  - "asymptotic advantage": $p^*_{AB} - p^*_B$
  - "integrated speedup":

$$\frac{1}{p^*_B - p^0_{AB}} \int_{p^0_{AB}}^{p^*_B} \text{speedup}(p)\,dp = \frac{1}{p^*_B - p^0_{AB}} \int_{p^0_{AB}}^{p^*_B} \frac{\text{lexp}_B(p)}{\text{lexp}_{AB}(p)}\,dp$$



There is Still a Problem

- The integrated speedup can still be infinite
- At p^0_{AB}, the speedup is usually infinite
  - although this occurs only at a point, it can cause the integral to diverge
  - example:



$$\int_1^2 \frac{(p-1)(b-a)+a}{(p-1)b} dp =$$

$$\int_1^2 \left[\frac{b-a}{b} + \frac{a}{b}\frac{1}{p-1}\right] dp =$$

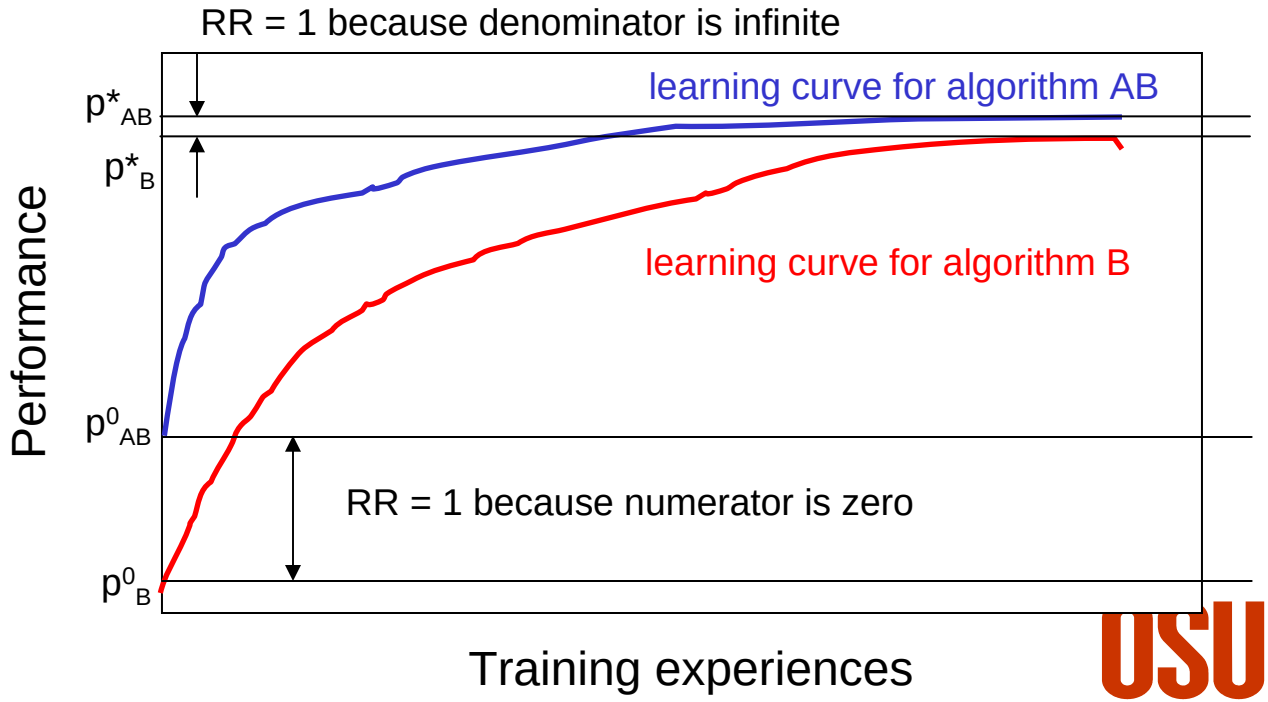
$$\frac{b-a}{b} + \frac{a}{b} \ln p \Big|_0^1 = \infty$$

Solution: Use Relative Reduction Instead

- "average relative reduction":

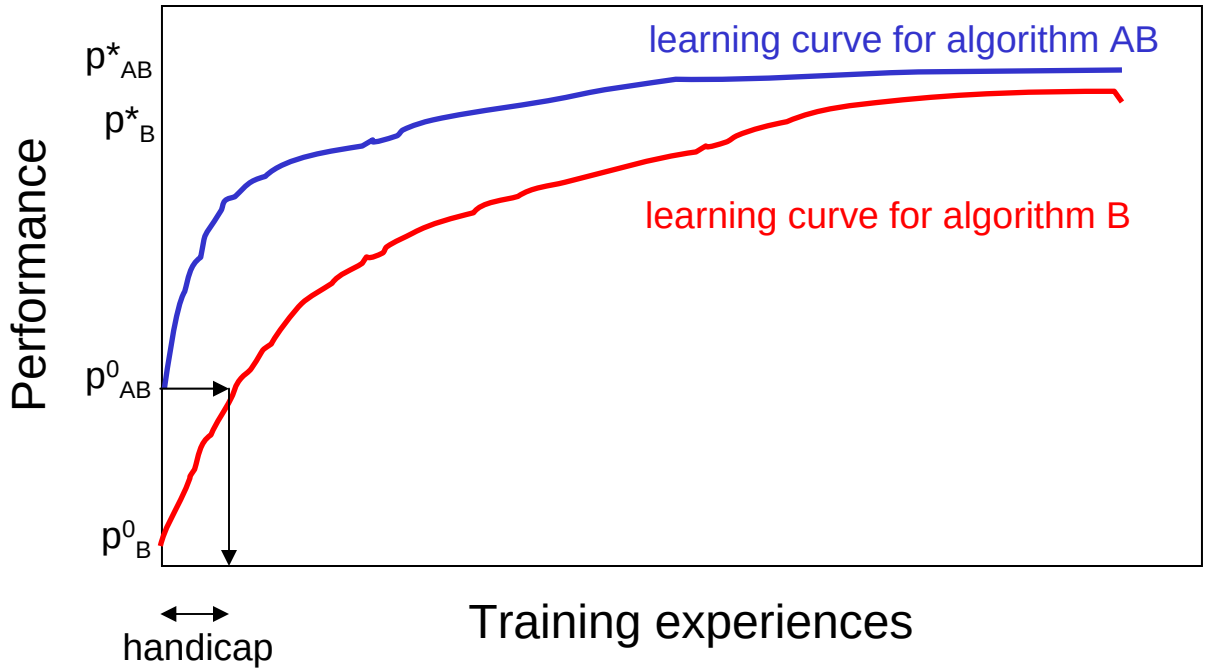
$$\frac{1}{p^*_{AB} - p^0_B} \int_{p^0_B}^{p^*_{AB}} \mathrm{RR}(p)dp = \frac{1}{p^*_{AB} - p^0_B} \int_{p^0_B}^{p^*_{AB}} 1 - \frac{\mathrm{lexp}_{AB}(p)}{\mathrm{lexp}_B(p)} dp$$

  - avoids (almost) all of the division-by-zero problems
  - define integral to run from $\min(p^0_B, p^0_{AB})$ to $\max(p^*_B, p^*_{AB})$



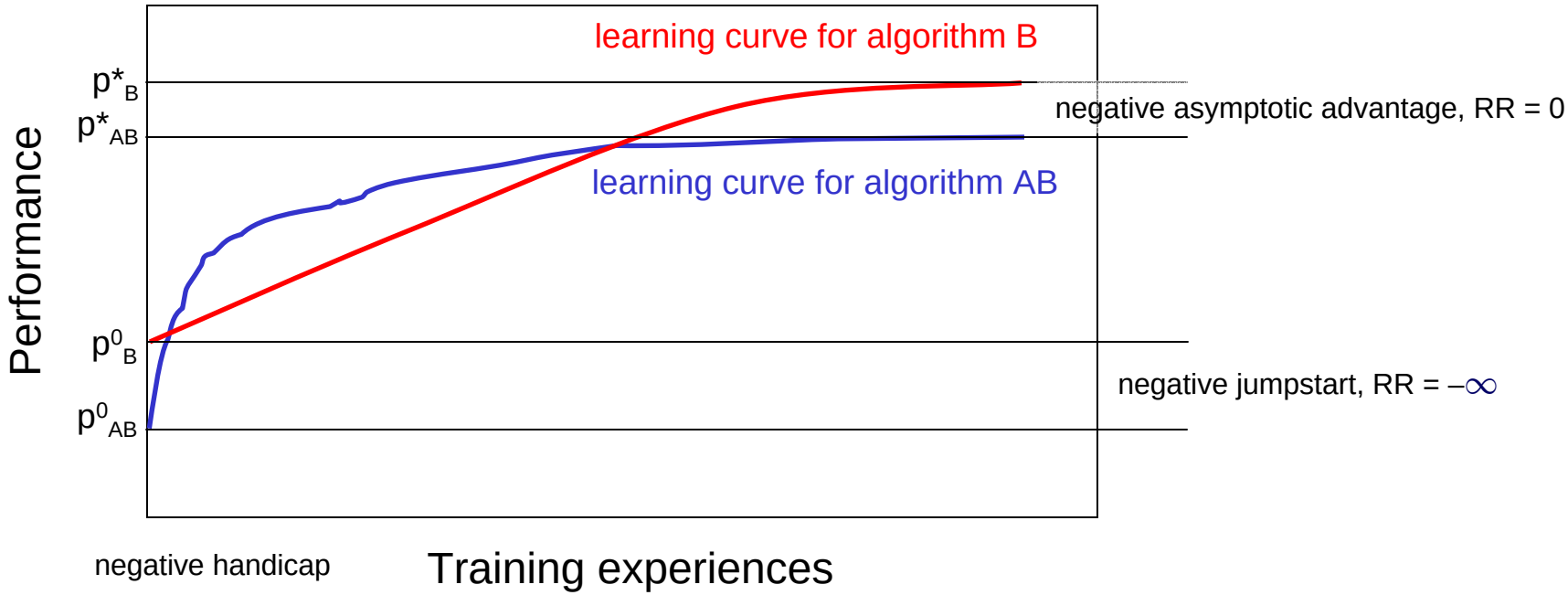
One More Quantity of Interest

- "handicap": How much training does it take B to reach performance level p^0_{AB}? (i.e., to overcome the jumpstart)



Corner Cases

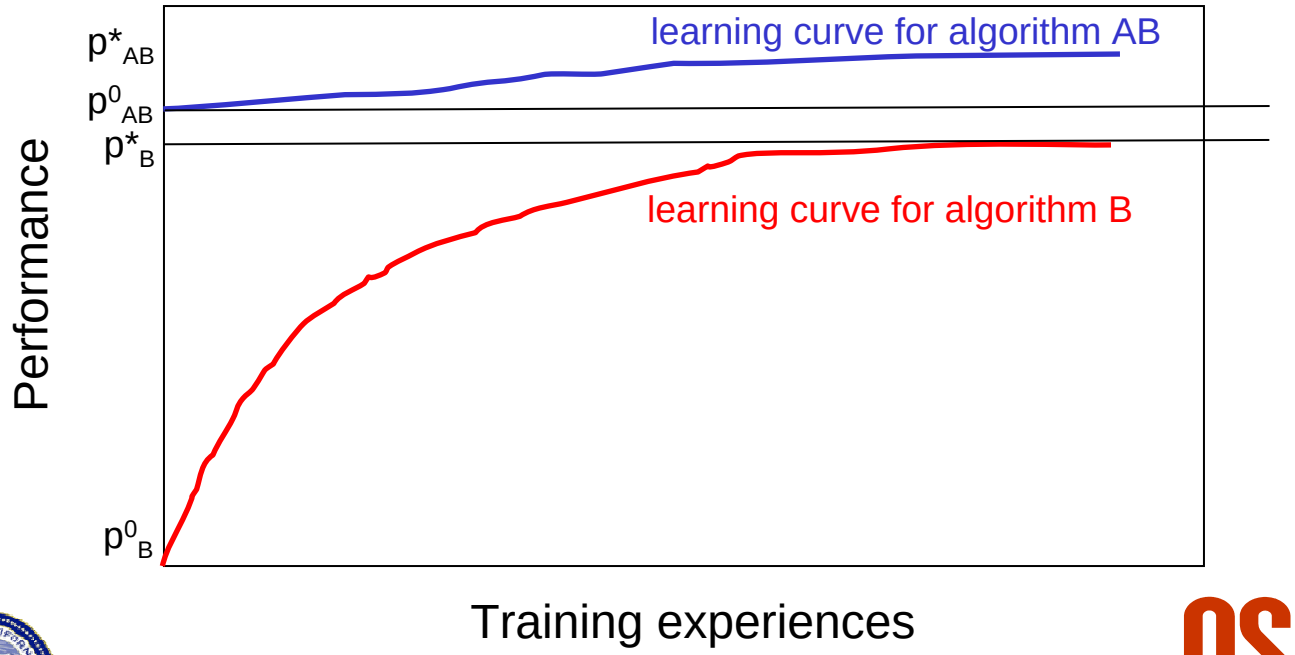
- Crossing curves: all 3 auxiliary measures can be negative



- Average RR will be $-\infty$. This will be rare, but it gives the no-transfer case the benefit.

Corner Cases (2)

- Transfer curve can be completely above the no-transfer curve:
  - large jumpstart
  - modest asymptotic advantage
  - infinite handicap
  - RR = 1 everywhere



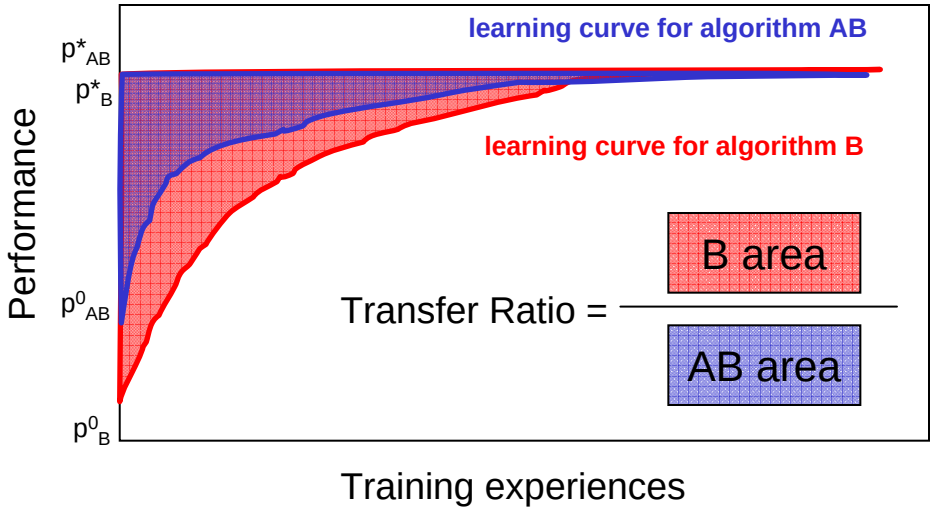
Additional Notes

- The algorithm(s) can attain performance level p multiple times. lexp(p) must be defined as the minimum of such occurrences
- The RR integral can be < 0, which indicates that transfer learning was worse than no-transfer
- In practice, each learning curve must be approximated, and because it is measured at fixed training experiences N_1, N_2, …, it is unlikely that the measured performance will exactly match a given target level p. Hence we propose approximating the learning curve by a piece-wise linear curve and interpolating p and N linearly to compute lexp(p). With this approximation, the integral can be computed exactly (over a series of trapezoids defined by perf(N_1), perf(N_2), …).
- Of course the measured learning curve is only an estimate of the true curve. We could generate bootstrap learning curves (via bootstrap replicates of the test data), compute the metrics for each curve, and then compute and report confidence intervals for the metrics.

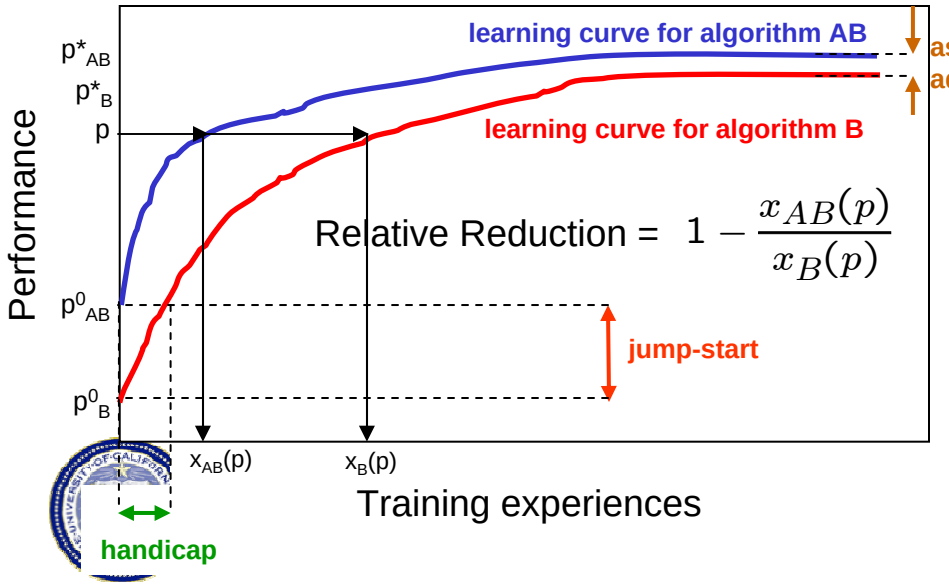
Summary

- Four proposed metrics:
  - average relative reduction in training time (sample size, number of training experiences)
  - jumpstart (initial advantage of transfer algorithm)
  - handicap (how long it takes the no-transfer algorithm to overcome the jumpstart)
  - asymptotic advantage (how much better the transfer learning algorithm does in the limit of large sample sizes)

Relative Reduction vs. Transfer Ratio



- Problems:
  1. Does not give instantaneous measure
  2. Assumes $p^*_{AB} = p^*_B$ (gives infinite ratio otherwise)
  3. Hard to explain (unclear units)
- Solution:
  1. Compute instantaneous measure, then integrate
  2. Integrate along the performance axis
  3. Measure reduction in amount of training needed ("amount reduced by 75%")



- Average relative reduction:

$$\frac{1}{p^*_{AB} - p^0_B} \int_{p^0_B}^{p^*_{AB}} 1 - \frac{x_{AB}(p)}{x_B(p)} dp$$

- Auxiliary metrics:
  - jump start: $p^0_{AB} - p^0_B$
  - asymptotic advantage: $p^*_{AB} - p^*_B$
  - handicap: $x_B(p^0_{AB})$