

Derivation: Scaling Exponent and Intercept Errors in Chinchilla Approach 2

We derive a closed-form expression for the vertex shift δw as a function of the loss surface parameters (α, β) and grid specification (W, n) .

1. Setup

The loss along an IsoFLOP contour is:

$$L(N; C) = E + AN^{-\alpha} + B \cdot 6^\beta C^{-\beta} N^\beta$$

Let $w = \log_{10}(N/N^*)$, where N^* is the true optimum. Then:

$$L(w) = E + P \cdot 10^{-\alpha w} + R \cdot 10^{\beta w}$$

where $P = A \cdot (N^*)^{-\alpha}$ and $R = B \cdot 6^\beta C^{-\beta} (N^*)^\beta$.

At the true optimum, the first-order condition gives:

$$\alpha P = \beta R$$

2. Parabola Fit via Least Squares

We sample at n equally-spaced points $w_i \in [-W, W]$ and fit a parabola:

$$\hat{L}(w) = a_0 + a_1 w + a_2 w^2$$

For symmetric grid points, the least-squares coefficients are:

$$a_1 = \frac{\sum_i w_i \cdot L(w_i)}{\sum_i w_i^2}$$
$$a_2 = \frac{n \sum_i w_i^2 L(w_i) - (\sum_i w_i^2)(\sum_i L(w_i))}{n \sum_i w_i^4 - (\sum_i w_i^2)^2}$$

3. Key Simplification: Normalize by R

Since $\alpha P = \beta R$, we have $P = (\beta/\alpha)R$. The loss becomes:

$$L(w) = E + R \left[\frac{\beta}{\alpha} \cdot 10^{-\alpha w} + 10^{\beta w} \right]$$

Define the **normalized loss contribution**:

$$\tilde{L}(w) = \frac{\beta}{\alpha} \cdot 10^{-\alpha w} + 10^{\beta w}$$

The actual loss is $L(w) = E + R \cdot \tilde{L}(w)$.

Critical observation: The constant E doesn't affect the parabola vertex (it shifts the parabola vertically but not horizontally). The scale factor R cancels in the ratio a_1/a_2 . Therefore:

$$\delta w = -\frac{a_1}{2a_2} = f(\alpha, \beta, W, n)$$

The vertex shift depends only on α , β , W , and n — not on C , E , A , B , or R .

4. Closed-Form Expression for δw

Substituting $\tilde{L}(w_i)$ into the least-squares formulas:

Vertex shift formula:

$$\delta w = -\frac{a_1}{2a_2} = -\frac{\sum_i w_i \cdot \tilde{L}(w_i)}{2 \sum_i w_i^2} \cdot \frac{n \sum_i w_i^4 - (\sum_i w_i^2)^2}{n \sum_i w_i^2 \tilde{L}(w_i) - (\sum_i w_i^2)(\sum_i \tilde{L}(w_i))}$$

where:

$$\tilde{L}(w) = \frac{\beta}{\alpha} \cdot 10^{-\alpha w} + 10^{\beta w}$$

and the sums are over equally-spaced points $w_i \in [-W, W]$.

Special Case: $\alpha = \beta$

When $\alpha = \beta$, the normalized loss simplifies to:

$$\tilde{L}(w) = 10^{-\alpha w} + 10^{\alpha w} = 2 \cosh(\alpha w \ln 10)$$

This is symmetric about $w = 0$, so $\sum_i w_i \tilde{L}(w_i) = 0$ (odd function times even function summed over symmetric points).

Therefore $a_1 = 0$ and $\delta w = 0$.

5. Intercept Error

The inferred optimum is:

$$\hat{N}^* = N^* \cdot 10^{\delta w}$$

Since the true scaling is $N^* = a_0 \cdot C^a$, the inferred scaling is:

$$\hat{N}^* = a_0 \cdot 10^{\delta w} \cdot C^a = \hat{a}_0 \cdot C^a$$

Intercept error:

$$\frac{\hat{a}_0 - a_0}{a_0} = 10^{\delta w} - 1$$

where $\delta w = f(\alpha, \beta, W, n)$ is the vertex shift formula above.

6. Exponent Preservation

Since δw is independent of C :

$$\log_{10}(\hat{N}^*) = \log_{10}(a_0) + \delta w + a \cdot \log_{10}(C)$$

Fitting $\log_{10}(\hat{N}^*)$ vs $\log_{10}(C)$ across multiple compute budgets:

- **Slope** = a (exact, because δw is constant)
- **Intercept** = $\log_{10}(a_0) + \delta w$ (shifted by δw)

Main Result:

- **Exponent error** = 0 (exactly, for any α, β, W, n)
- **Intercept error** = $10^{\delta w(\alpha, \beta, W, n)} - 1$

7. Numerical Examples

For $n = 15$ points:

| SURFACE | A | B | W | ΔW | INTERCEPT ERROR |
|----------------|-------|-------|-----|------------|-----------------|
| Symmetric | 0.31 | 0.31 | 1.0 | 0 | 0% |
| Chinchilla | 0.34 | 0.28 | 0.3 | 0.0014 | 0.33% |
| Chinchilla | 0.34 | 0.28 | 1.0 | 0.0157 | 3.7% |
| Chinchilla | 0.34 | 0.28 | 2.0 | 0.0626 | 15.5% |
| High imbalance | 0.465 | 0.155 | 1.0 | 0.0795 | 20.1% |
| High imbalance | 0.465 | 0.155 | 2.0 | 0.2992 | 99.2% |

8. Summary

The vertex shift δw is a closed-form function of α, β, W , and n :

$$\delta w = f(\alpha, \beta, W, n)$$

Key properties:

- $\delta w = 0$ when $\alpha = \beta$ (symmetric loss)
- δw grows with $|\alpha - \beta|$ (asymmetry)
- δw grows with W (wider sampling range)
- δw is **independent of C** \rightarrow exponent is exactly preserved