The µP Cheat Sheet

What is a parametrization?

A parametrization is a set of rules that determine the values of three types of hyperparameter (HP) in a model:

1. parameter multipliers (α)
2. initialization standard deviations (σ)
3. learning rates (η)

such that

for width-parametrizations

$$W_t = \alpha_W \cdot w_t$$
$$w_0 = \mathcal{N}(0, \sigma_W^2)$$
$$w_{t+1} = w_t - \eta_W \cdot [\text{update}]$$

for depth-parametrizations

$$\text{residual-block}(x) = \alpha_{\text{res}} \cdot f_{\text{res}}(x) + x$$

and $\sigma_{\text{res}}, \eta_{\text{res}}$ apply to all W on the residual branch

These rules are typically functions of the width (d) and/or layers (ℓ) of the model.

Rather than defining rules for individual Ws, parameterizations tend to define rules according to three tensor types:

tensor type	fan-in(W)	fan-out(W)	example
1. Input	$\Theta(1)$	$\Theta(d)$	encoder, embedding, bias, norm params
2. Hidden	$\Theta(d)$	$\Theta(d)$	linear layer, convolution
3. Output	$\Theta(d)$	$\Theta(1)$	decoder, readout
(Residual)			(any parameter tensor on a residual branch)

Definitions of key parametrizations

The following table-pair captures important width-parametrizations from the literature. To derive a parametrization, select its column from the left table and take the entries from the right table of the corresponding color:

Param. feature	SP	NTP-na	NTP-fa	µP-na	µP	u-µP
down-scale η_{in}	**n/a**	✗	✗	✗	✗	✓
'full-alignment'	**n/a**	✗	✓	✗	✓	✓
down-scale σ_{out}	✗	✗	✗	✓	✓	✓

HP	tensor type					
	input		hidden		output	
σ	1		$1/\sqrt{d}$		$1/\sqrt{d}$	$1/d$
(Adam) η	1	$1/\sqrt{d}$	$1/\sqrt{d}$	$1/d$	$1/\sqrt{d}$	$1/d$

No values are given for αs due to **abc-symmetry**, which states that models are invariant to changes of α, σ, η of the form $\alpha_W \times\!= \theta, \sigma_W \div\!= \theta, \eta_W \div\!= \theta$ for a given W, θ. Different values of θ define different parametrizations in the same *equivalence class*. θ is always chosen above such that $\alpha = 1$ for the sake of comparison, though other forms are used in the literature (e.g. the Mean Field Parametrization [todo] is in the same equivalence class as µP, but usually presented differently). The only parametrization which specifies a preferred form is u-µP, which uses $\sigma = 1$ for numerical stability. Also note that:

- σ, η are only *proportional* to the values in the right table—the constant of proportionality is a tunable HP.
- The above ηs apply to all **optimizers** which guarantee $\Theta(1)$-sized updates (e.g. Adam, Shampoo, Muon). Table 1 of [todo] shows adjustments for other optimizers.
- Standard Parametrization (**SP**)'s rules for η are presented inconsistently in the literature. Hence they are dropped entirely here.
- **'fa/na'** denotes full/no-alignment assumptions from [todo]
- The typical presentation of Neural Tangent Parametrization (**NTP**) is different to the one shown here (see Table 1 [tp4]) and has been shown to scale poorly. The 'fa/na' variants use more appropriate ηs.
- **u-µP**'s down-scaled η_{in} has only been validated on embedding-style input layers.

TODO: something on depth

F.A.Q.

What are parametrizations trying to do?

This is some text that will appear in the first column.

What parametrization should I use?

This is text for the second column.

Any practical tips for applying this?

This is text for the third column.

Charlie Blake, 2025 github.com/thecharlieblake/mup-cheat-sheet thecharlieblake @ gmail.com