

gradient variance  $\propto \frac{1}{\text{batch size}}$ ; lower variance  $\Rightarrow$  larger learning rate

$(\text{batch size})_{\text{new}}$   
 $= \text{BS} + \text{BS} \times k$

$\Rightarrow$

$(\text{learning rate})_{\text{new}}$   
 $= \text{LR} + \text{LR} \times k$

$\text{model capacity} \uparrow$   
 $\text{small dataset} \Rightarrow$

$\Rightarrow$

$\text{regularisation} \uparrow$   
 $\text{aggressive regula.}$

$\times \| \theta \|^2$

Optimisation: find optimal parameters to minimise predefined loss func.

Regularisation: add a penalty term to loss func during optimisation to prevent overfitting

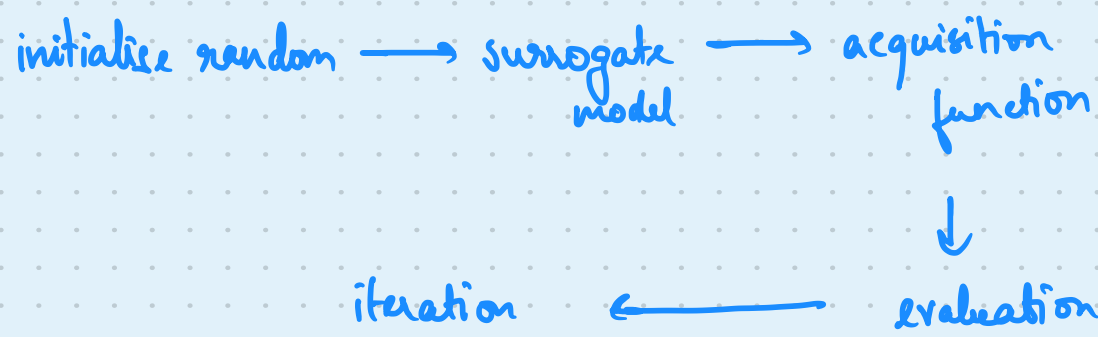
# Hyperparameter Optimisation

## No Free Lunch Theorem

$$\sum_P E_A(P) = \sum_P E_B(P)$$

where  $P$ : set of all problems

## Bayesian/Sequential Model Based Optimisation



Say we use a Gaussian Process model for surrogate

We use a Gaussian Process (GP) prior.

- A GP gives us a predictive distribution at any point  $x$ :

$$M_t(x) \sim \mathcal{N}(\mu_t(x), \sigma_t^2(x))$$

where

- $\mu_t(x)$  is the predicted mean (our best guess).
- $\sigma_t(x)$  is the uncertainty.

balance exploitation and exploration

Common acquisition function: Expected Improvement

$$EI_{y^*}(x) = \int_{-\infty}^{\infty} \max(y^* - y, 0) P_m(y|x) dy$$

$\downarrow$  min so far

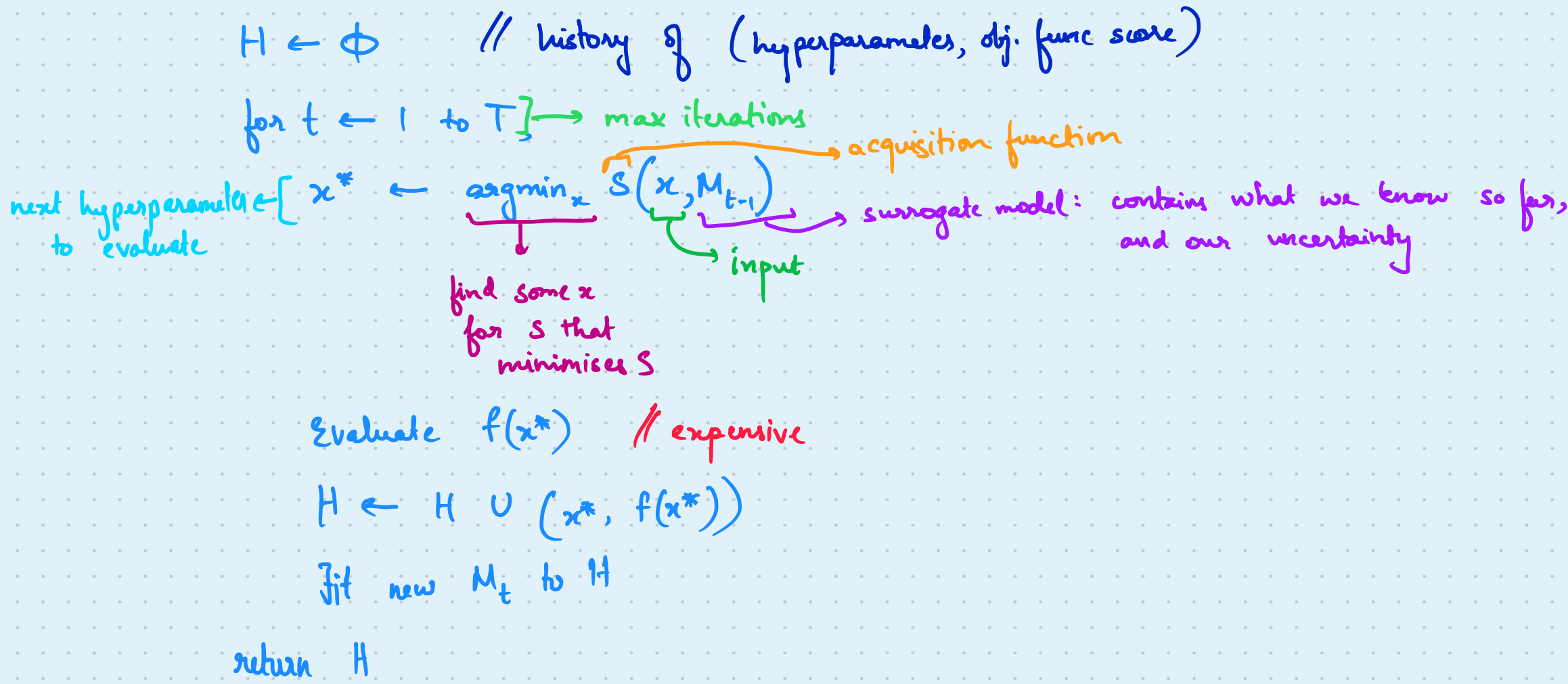
$\uparrow$  new obj. function score

$\rightarrow$  hyperparameter

$\leftarrow$  surrogate model

## SMBO Algorithm

SMBO( $f, M_0, T, S$ )



## Hyperparameters in common ML models

Decision tree: max depth, leaves per split

K-NN, K-means:  $k$

Random Forest: no. of trees

Naive Bayes: smoothing parameters  $\rightarrow$  probabilities multiplied, small constant to prevent zeroes when data not present in training

$O(W^3)$

Logistic Regression: solver

SVM: kernel  $\begin{cases} \text{linear} \\ \text{poly} \\ \text{Radial Basis Function (RBF)} \end{cases}$

Hidden Markov: epochs, length of observations, hidden states

Gaussian Mixture: no. of Gaussians

Neural networks: learning rate, no. of epochs, batch size, step

$\rightarrow$  decay: time-based, step-based, cosine

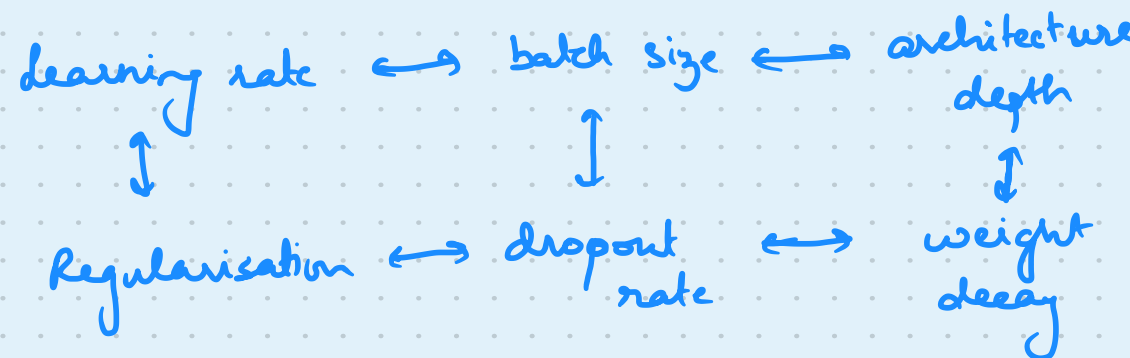
$\rightarrow$  adaptive: ADAM, Adagrad, RMSprop

$\rightarrow$  no. of batches processed in one epoch

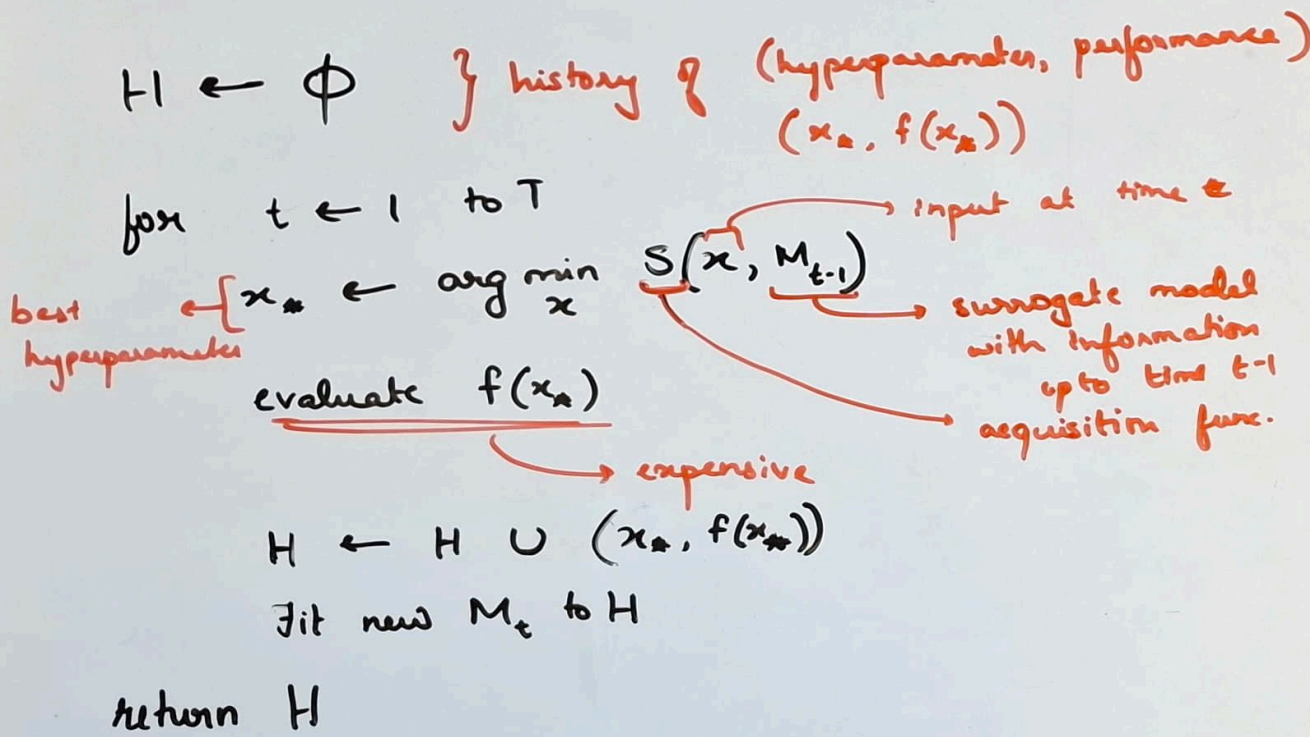
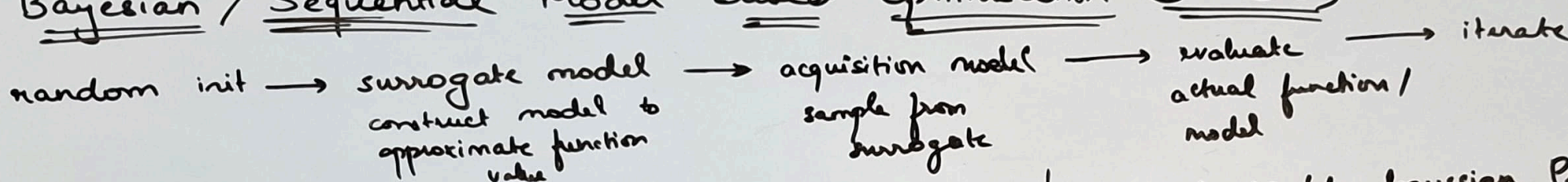
step = forward + backward + update

## Hyperparameters Interaction Webs

Nothing exists in isolation



## Bayesian/Sequential Model Based Optimisation (SMBO)



## Surrogate Model: Gaussian Process

$$f(x_n) | D_t \sim \mathcal{N}(\mu_t(x_n), \sigma_t^2(x_n))$$

$\leftarrow$  observed data until  $t$

$\leftarrow$  best guess so far

$\leftarrow$  uncertainty

## Acquisition Function: Expected Improvement

$$EI_{y^*}(x) = \int_{-\infty}^{\infty} \max(y^* - y, 0) P_m(y|x) dy$$

$\leftarrow$  minimum so far

$\leftarrow$  surrogate model