(a)

i. The parameter m will keep the track of last direction of the gradient descent, and helps to counteract the current part of direction which not the true direction, and strengthen the true direction of the gradient descent.

ii. Positions which the computing gradients are low will get corresponding larger updates.

Reasons about helping: Since the gradients calculated are different, some of the positions are higher, some of are lower. In this way, the learning rate is self adjusted by the gradients, and the model will converge faster.

(b)

i.

$$\mathbb{E}_{p_{\text{drop}}}\ [\mathbf{h}_{drop}]_i = h_i \implies \gamma(1-p_{drop})h_i = h_i \implies \gamma = \frac{1}{1-p_{drop}}$$

ii.

Training: As a regulation method, generalize the model
Testing: Since the droping procedue is random, this will cause different evaluation results when you perform multiple testing.