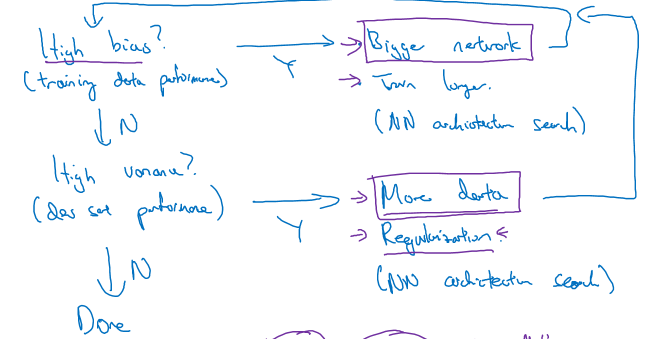
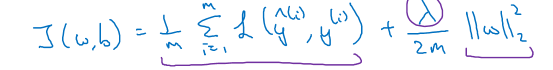
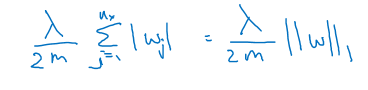
# 训练集/验证集/测试集数据的划分

1. 数据量小的时候：70/30, 或 60/20/20
2. 数据量大的时候：只需要验证集和测试集数量足够大即可
3. 在数据分布上，验证集/测试集要一样，但是训练集不一定要一样

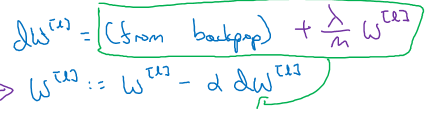


# 正则化(Regularization) – reduce overfitting



**L1正则化**：

**L2正则化**：



**Dropout 正则化**：随机临时性的屏蔽某些内部单元

**其他**：data等

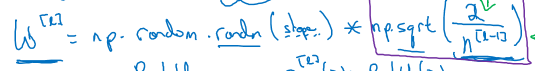
# Normalizing

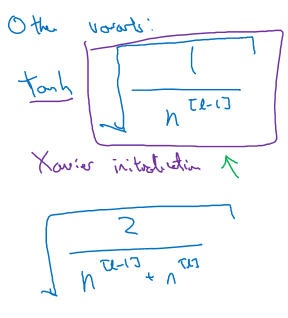
# Vanishing/exploding gradients

当NN层数很多时，w>I，梯度太大；w<I，梯度太小，需要在初始化时进行调节

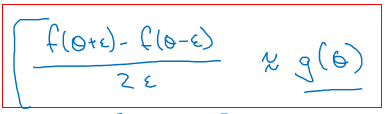


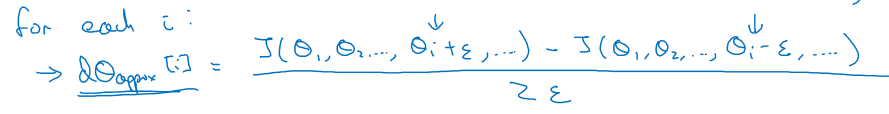
采用几种初始化方式避免这种情况：

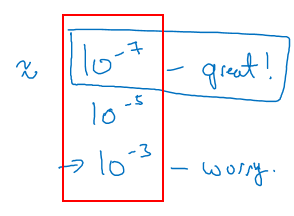




# 检查公式计算是否有bug







1. 耗计算时间，所以只是在debug时使用
2. 如果不满足条件，重点检查dw和db
3. 可以归一化，但是不支持dropout
4. 可以随机初始化

# Mini-batch梯度下降

对于小的训练集（m<2000）使用mini-batch梯度下降

典型的mini-batch size：64、128、256、512（2的n次方），注意要CPU/GPU配合

* Mini-batch size = m：batch gradient descent（BGD），每次迭代都训练所有的数据，耗时长，但是梯度下降摆动小，受噪音影响小
* Mini-batch size = 1：stochastic gradient descent（SGD），每次迭代只训练一个数据，下降迅速，但是摆动大，受噪音影响大

# 指数加权平均值（Exponentially weighted averages）--修正bias，缓和梯度下降摆动



大约涉及t、t-1、t-2、… 共1/(1-β)个时间点

## momentum

用在梯度下降中，缓和下降过程中的摆动：

On iteration 𝑡：

Compute 𝑑𝑊,𝑑𝑏 on the **current mini-batch：**

𝑣𝑑𝑊=𝛽𝑣𝑑𝑊+(1−𝛽)𝑑𝑊

𝑣𝑑𝑏=𝛽𝑣𝑑𝑏+(1−𝛽)𝑑𝑏

𝑊=𝑊−𝛼𝑣𝑑𝑊

𝑏=𝑏−𝛼𝑣𝑑𝑏

Hyperparameters: 𝛼,𝛽

实践中，**𝛽 =0.9**效果比较好，当然也可以尝试其他的值

## RMSprop

On iteration 𝑡：

Compute 𝑑𝑊,𝑑𝑏 on the **current mini-batch：**

S𝑑𝑊=𝛽2S𝑑𝑊+(1−𝛽2)𝑑𝑊2

S𝑑𝑏=𝛽2S𝑑𝑏+(1−𝛽2)𝑑𝑏2

𝑊=𝑊−𝛼

𝑏=𝑏−𝛼

Hyperparameters: 𝛼,𝛽2

ε=10-8

## Adam

On iteration 𝑡：

Compute 𝑑𝑊,𝑑𝑏 on the **current mini-batch：**

𝑣𝑑𝑊=𝛽1𝑣𝑑𝑊+(1−𝛽1)𝑑𝑊 --- momentum

𝑣𝑑𝑏=𝛽1𝑣𝑑𝑏+(1−𝛽1)𝑑𝑏

S𝑑𝑊=𝛽2S𝑑𝑊+(1−𝛽2)𝑑𝑊2 --- RMSprop

S𝑑𝑏=𝛽2S𝑑𝑏+(1−𝛽2)𝑑𝑏2

=

=

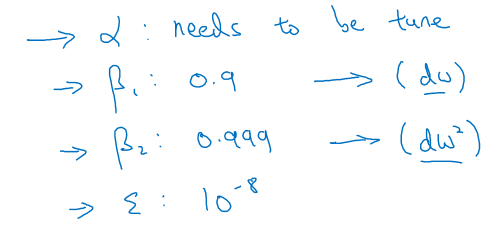
=

=

𝑊=𝑊−𝛼

𝑏=𝑏−𝛼

**Hyperparameters choice:**

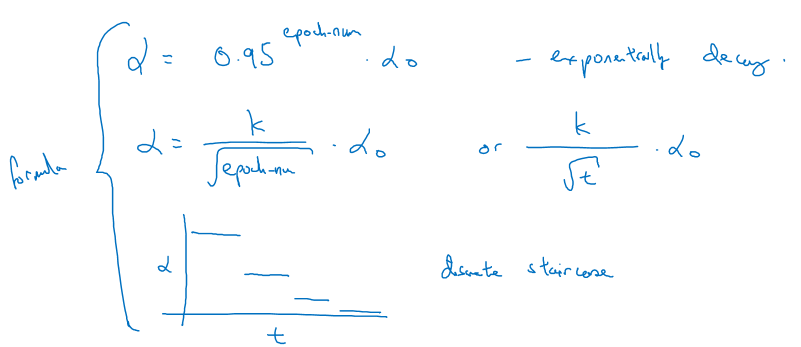
****

# 学习率（α）衰减

随着下降逐步缩小学习率。1 epoch指一段特定的时间，在这段时间中学习率是相同的。

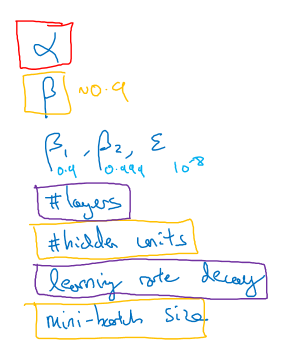
衰减率：decay-rate

其他公式：



也可以手工手工设定

# Hyperparameter调整过程



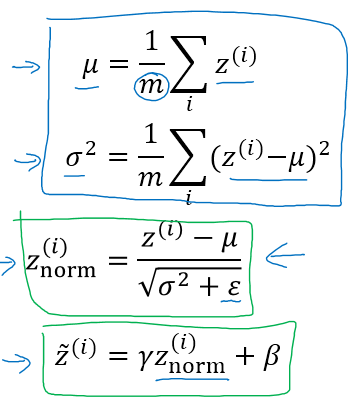
* **调整顺序：**
  + **红色框**是最重要，应该最优先调整
  + 橙色框次重要，第二批调整
  + 紫色，第三批调整
  + 其他的，可用缺省值，或者最后调整
* 不要使用grid方式
  + 传统的机器学习，使用grid，取最优值。适用于超参数比较少的情况
  + 深度学习中，因为不知道哪个超参数重要，所以一组超参数同时取随机值。
* 取随机值时，要使用合适的范围
  + 先在大区域中随机抽样，取效果比较好的区域，在此区域内提高密度再随机抽样
  + 在不同的区域，超参数的敏感度不一样，有时随机抽样需要变换尺度：线性、log、指数，等

# Batch Normalization

## Normalization输入 --- 加速学习过程

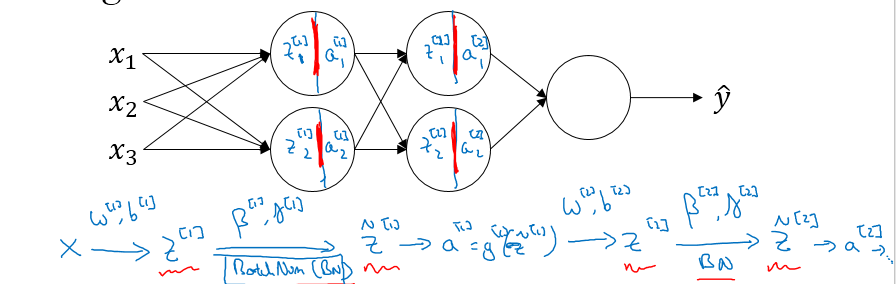
实践中，多数是对z做normalize而不是对a

Z[l](i)缩写为Z(i)=Z(1),Z(2),……,Z(m)



和：可以训练学习的Hyperparameter，用W、b、dW，db类似的方式计算

用代替 Z[l](i)：

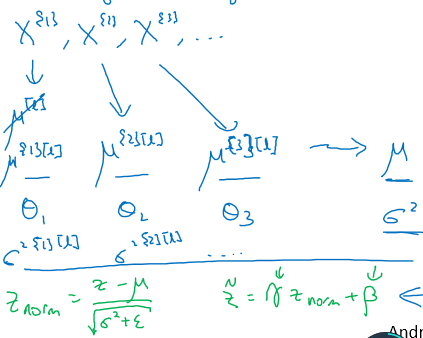


在mini-batch中使用时，注意：

* 每个μ和σ2只适用于这个mini-batch
* 会增加噪音
* 有少量的正则化效果

## 在测试集上使用Batch Norm

在测试集使用Batch Norm时，μ和σ2可使用跨mini-batch的指数加权平均值法来确定：



# 多分类问题

# 深度学习框架

选择深度学习框架考虑：

* 容易编程和发布
* 运算速度
* 真正的开源