统一梯度下降

普通训练通过计算一个batch中的平均梯度作为参数更新方向，这具有明显缺陷：

1. 易受恶性节点的影响，对梯度小的样本不公平。
2. 采用的代数平均过于简单。

统一梯度下降算法：

核心思想：

选取一个更新方向，使得大多数（尽量最多）样本误差在该方向上减少，从而避免了梯度冲突与“南辕北辙”，且以潜在的统一最小值为优化目标而不是单个目标的误差最小值为优化目标。集体重于个人。

实现方式一（随机寻找）：

以梯度平均值为初始值，添加微小扰动寻找更新值，使得更多样本误差在该方向上减少（点积为正）。

实现方式二（完备算法）：

从样本梯度中选取n个构成分割面，分割面的垂直向量即为更新方向。

复杂度： b为batch中的样本数，n为单个样本的维度

证明：

若最优分割面经过m个梯度向量，m<n，该超平面的自由度小于向量空间的维度，因此其有n-m个自由度可以自由变化，若不论如何变化都不经过n个梯度向量，则向量空间可以降维至n’,此时分割面经过了n’个梯度向量，证毕。

实现方式三（模糊数学求非最优解）：

在想。

实验验证：

1. 实验组1采用统一梯度下降，与对照组1（裸梯度下降）相比，收敛速度更快，准确度更高，效果明显。
2. 实验组2（添加了adam优化的动量算法和自适应学习率，以期减少adam的对比干扰）比对照组2（采用adam优化的普通的标准训练）更快收敛，准确度更高。

实验不足：

1. 因快速验证所需，仅在mnist数据集上实验，样本分布较简单，不能很好凸显出统一梯度下降在复杂分布的数据集上的潜力。
2. 统一梯度下降方向的计算方式不够好，复杂度较高。

其他参考：Trimmed Mean、Median of Gradients、Consensus Optimization等类似算法都和统一梯度下降算法不一样，它们计算的是修剪过的平均值或中位数，且是在分布式训练上的算法。