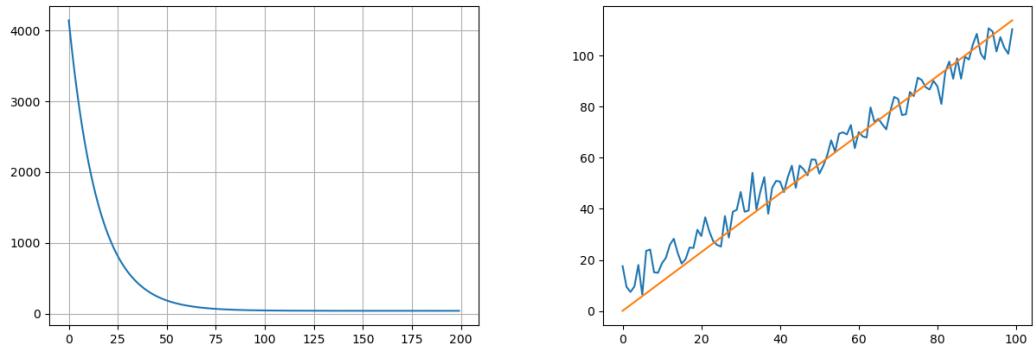
1 BGD, SGD 和 MBGD 对训练的影响

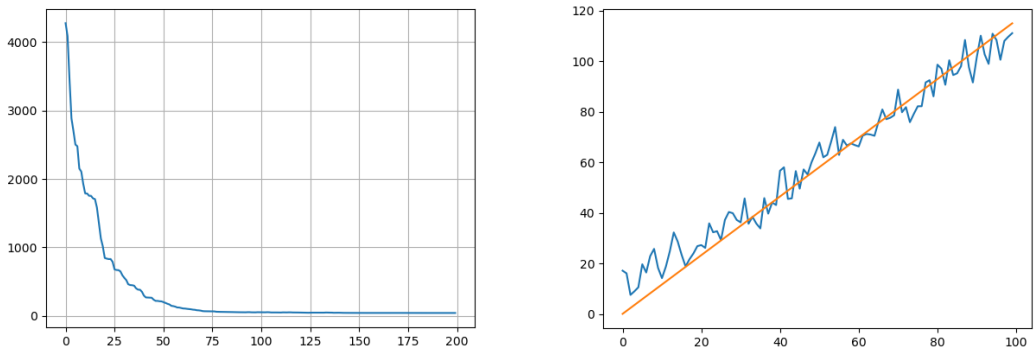
对于三种梯度下降算法，共同采用以下参数（MBGD 中 batch_size=10）：

迭代次数 epochs	200
学习率 lr	1e-5

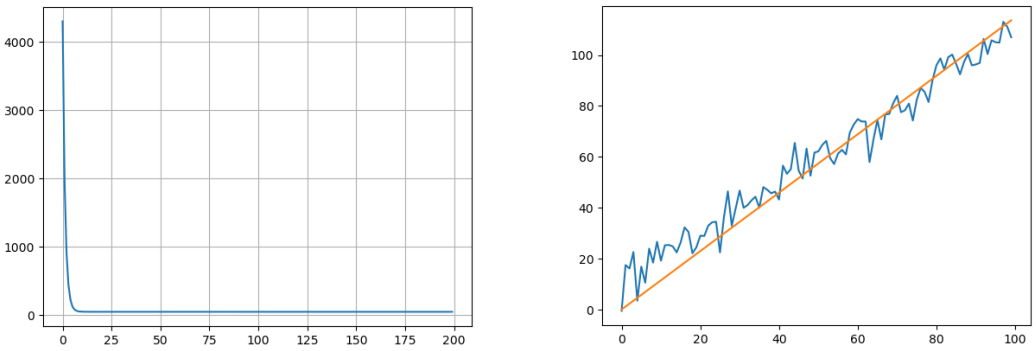
初始权重在 0 附近，不使用 normalization，训练结果如下：



BGD 的 loss 曲线和拟合曲线



SGD 的 loss 曲线和拟合曲线



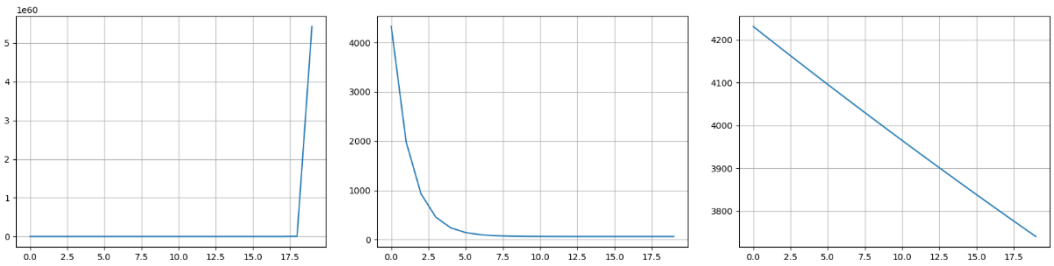
MBGD 的 loss 曲线和拟合曲线

可以发现，BGD 和 SGD 均在训练到 100 轮左右后，loss 曲线几乎不再下降，

达到了良好的训练效果。然则在这 100 轮中，BGD 遍历了 100 次数据集，而 SGD 仅访问了 100 个样本，相当于仅遍历一次数据集。因此，可以推断 SGD 的训练速度比 BGD 更快。MBGD 在训练 10 轮左右时达到良好的效果，此时该方法遍历了 10 次数据集，因此速度介于 BGD 和 SGD 之间。

除此之外，通过观察 loss 曲线的变化，容易发现 SGD 的 loss 曲线在下降过程中出现的波动会明显强于另外两种方法，这说明 SGD 在每一次更新权重时易受到噪声的影响，而另外两种方法每次更新能更准确地指向极值方向。

适当调整学习率，发现学习率的改变对三种算法的影响相同：学习率增大到 1e-3，三种算法均不收敛；减小学习率，训练速度变慢：



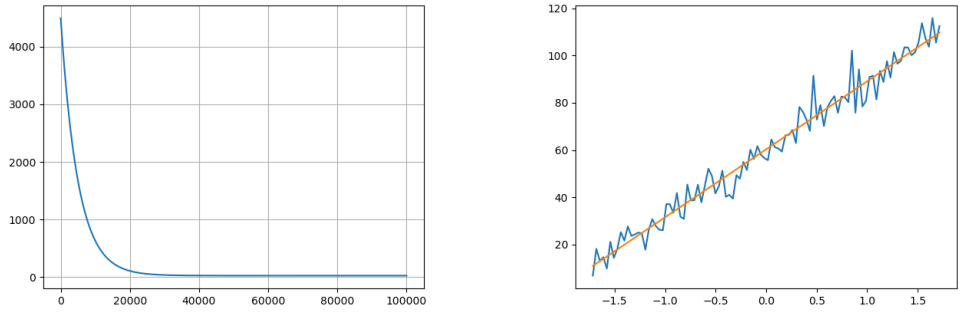
MBGD 的 loss 曲线，epochs=20，左 lr=1e-3，中 lr=1e-5，右 lr=1e-7

总结如下：

1. SGD 的训练速度最快，MBGD 次之，BGD 最慢；
2. 更新权重时，SGD 易受噪声影响，另外两种方法能更准确指向极值方向；
3. 过大的学习率导致溢出，过小的学习率导致训练速度变慢。

2 归一化对训练的影响

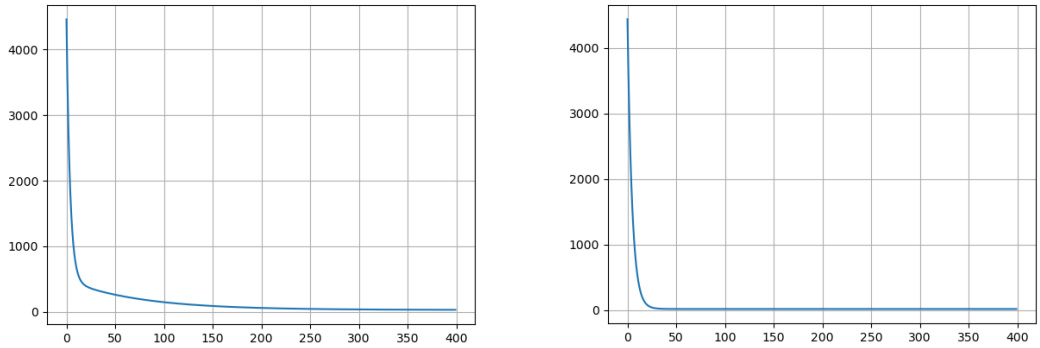
初始权重在 0 附近，分别使用两种 normalization，可以发现当学习率仍为 1e-5 时，训练速度非常缓慢：



MBGD 和 mean normalization，lr=1e-5

将适当增大学习率，两种 normalization 的训练速度都得到明显提升，以 MBGD 为例，设置参数如下：

迭代次数 epochs	400
学习率 lr	1e-2
batch_size	10



lr=1e-2，**左图** min-max normalization，**右图** mean normalization

可以观察到，min-max normalization 在 300 轮左右到达最低点，mean normalization 在不到 50 轮时就能到达最低点。在该实验条件下，mean normalization 的收敛速度明显快于 min-max normalization。若不使用 normalization，增大学习率到 1e-3 附近，算法不收敛。可以说明，normalization 和学习率共同影响训练效果。在该实验中，使用较大的学习率时，normalization 会有利于算法的收敛；使用较小的学习率时，normalization 会使收敛变慢。