Adadelta

下面我们实现一下 Adadelta

```
def adadelta(parameters, sqrs, deltas, rho):
    eps = 1e-6
    for param, sqr, delta in zip(parameters, sqrs, deltas):
        sqr[:] = rho * sqr + (1 - rho) * param.grad.data ** 2
        cur_delta = torch.sqrt(delta + eps) / torch.sqrt(sqr + eps) * param.grad.data
        delta[:] = rho * delta + (1 - rho) * cur_delta ** 2
        param.data = param.data - cur_delta
```

```
import numpy as np
import torch
from torchvision.datasets import MNIST # 导入 pytorch 内置的 mnist 数据
from torch.utils.data import DataLoader
from torch import nn
from torch.autograd import Variable
import time
import matplotlib.pyplot as plt
%matplotlib inline

def data_tf(x):
    x = np.array(x, dtype='float32') / 255
    x = (x - 0.5) / 0.5 # 标准化，这个技巧之后会讲到
    x = x.reshape((-1,)) # 拉平
    x = torch.from_numpy(x)
    return x

train_set = MNIST('./data', train=True, transform=data_tf, download=True) # 载入数据
集，申明定义的数据变换
test_set = MNIST('./data', train=False, transform=data_tf, download=True)

# 定义 loss 函数
criterion = nn.CrossEntropyLoss()
```

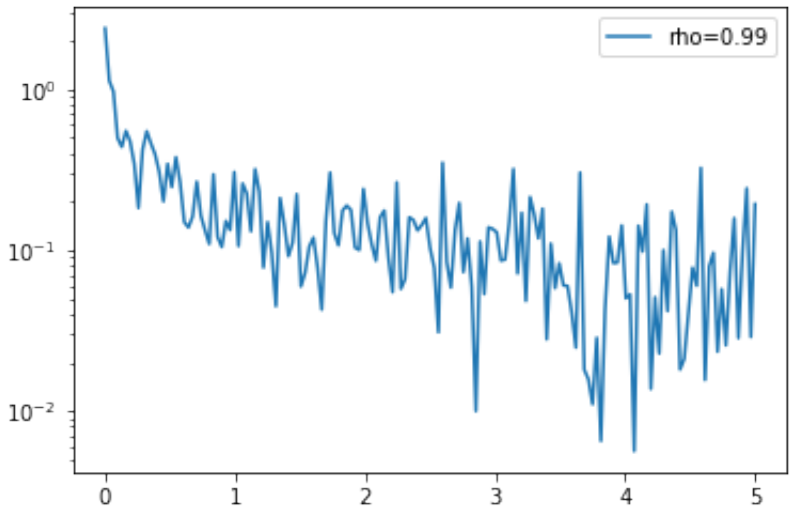
```
train_data = DataLoader(train_set, batch_size=64, shuffle=True)
# 使用 Sequential 定义 3 层神经网络
net = nn.Sequential(
    nn.Linear(784, 200),
    nn.ReLU(),
```

```
    nn.Linear(200, 10),
)

# 初始化梯度平方项和 delta 项
sqrs = []
deltas = []
for param in net.parameters():
    sqrs.append(torch.zeros_like(param.data))
    deltas.append(torch.zeros_like(param.data))

# 开始训练
losses = []
idx = 0
start = time.time() # 记时开始
for e in range(5):
    train_loss = 0
    for im, label in train_data:
        im = Variable(im)
        label = Variable(label)
        # 前向传播
        out = net(im)
        loss = criterion(out, label)
        # 反向传播
        net.zero_grad()
        loss.backward()
        adadelta(net.parameters(), sqrs, deltas, 0.9) # rho 设置为 0.9
        # 记录误差
        train_loss += loss.data[0]
        if idx % 30 == 0:
            losses.append(loss.data[0])
        idx += 1
    print('epoch: {}, Train Loss: {:.6f}'
          .format(e, train_loss / len(train_data)))
end = time.time() # 计时结束
print('使用时间: {:.5f} s'.format(end - start))
```

```
epoch: 0, Train Loss: 0.365601
epoch: 1, Train Loss: 0.159966
epoch: 2, Train Loss: 0.123347
epoch: 3, Train Loss: 0.102201
epoch: 4, Train Loss: 0.087986
使用时间: 59.26491 s
```

```
x_axis = np.linspace(0, 5, len(losses), endpoint=True)
plt.semilogy(x_axis, losses, label='rho=0.99')
plt.legend(loc='best')
```

```
<matplotlib.legend.Legend at 0x103f3a5f8>
```



可以看到使用 adadelta 跑 5 次能够得到更小的 loss

小练习：思考一下为什么 Adadelta 没有学习率这个参数，它是被什么代替了

当然 pytorch 也内置了 adadelta 的方法，非常简单，只需要调用 `torch.optim.Adadelta()` 就可以了，下面是例子

```
train_data = DataLoader(train_set, batch_size=64, shuffle=True)
# 使用 Sequential 定义 3 层神经网络
net = nn.Sequential(
    nn.Linear(784, 200),
    nn.ReLU(),
    nn.Linear(200, 10),
)

optimizer = torch.optim.Adadelta(net.parameters(), rho=0.9)

# 开始训练
start = time.time() # 记时开始
for e in range(5):
    train_loss = 0
```

```
    for im, label in train_data:
        im = Variable(im)
        label = Variable(label)
        # 前向传播
        out = net(im)
        loss = criterion(out, label)
        # 反向传播
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
        # 记录误差
        train_loss += loss.data[0]
    print('epoch: {}, Train Loss: {:.6f}'
          .format(e, train_loss / len(train_data)))
end = time.time() # 计时结束
print('使用时间: {:.5f} s'.format(end - start))
```

```
epoch: 0, Train Loss: 0.356505
epoch: 1, Train Loss: 0.158333
epoch: 2, Train Loss: 0.120510
epoch: 3, Train Loss: 0.100807
epoch: 4, Train Loss: 0.084741
使用时间: 47.90947 s
```

小练习：看看 pytorch 中的 adadelta，里面是有学习率这个参数，但是前面我们讲过 adadelta 不用设置学习率，看看这个学习率到底是干嘛的