SeqDGM 和 SeqDGM-A 的采样方法补充材料

很多推导我直接省略了，有可能的话请推导一遍验证一下. Theano 相关的部分我昨天验证过，最好也试下.

1. 首先将模型里的参数整理一遍:
   a. 模型中的不变量.
      句子序列, label, mask, 各种 dropout 等层中的随机数以及采样变分变量时候用的随机数 eps. 和不变量对应的是需要被求导的参数.
   b. Classisifer.
      分类器和模型的其他部分会共用一些参数: 在基础的 M1+M2 模型中, 共用了词向量参数 e; 在 AuxiliaryDGM 中, 共用的部分变多, 包含了变分变量 a 之前计算图中的所有节点(多了 Encoder 部分和 a 的采样部分), 当然还是可以用 e 表示.
      分类器自身也有独立的参数, 如 LSTM 中的参数. 这部分用 w 表示.
   c. Inference 和 Generation.
      分别用\phi 和\theta 去表示和分类器不共用的参数. \phi 等于变分变量 h 之前计算图中的节点减去 e. 注意到 h=F(x,a,label)或者(x,label), label 实际上是不变量. 因此在 M1+M2 模型中 Inference 和 Classifier 只共用了词向量 e; 在 AuxiliaryDGM 中共用情况见上节. 剩余不共用的参数构成了\phi
      \theta 则较为简单, Generation 参数排除词向量.

2. 主要参数可以用\phi \theta e w 来表示. 设 Label 数据为D_l, 大小为N_l, Unlabel 数据为D_u, 大小为N_u. 总的损失函数为:

$$C = \frac{1}{N_l+N_u}\left\{\sum_{<x,y>\in D_l} L(x,y;\phi,\theta,e) + \alpha(-\log q(y|x;w,e)) + \sum_{x\in D_u} U(x;\phi,\theta,w,e)\right\}$$

其中:

$$L(x,y;\phi,\theta,e)$$

$$= -E_{q(a,h|x,y;\phi,e)}[\log p(x|h,a,y;\theta,e)] - E_{q(a,h|x,y;\phi,e)}\left[\log\frac{p(a|y,h;\theta)p(h)}{q(a,h|x,y;\phi,e)}\right]$$

第一项重构误差用 MC 采样方法求, 第二项为广义上的 KL 距离, 用解析式求, 不再详细叙述.

实际上可以将重构误差减去一个 baseline 不影响所有参数的梯度方向:

$$\text{重构误差} = -E_{q(a,h|x,y;\phi,e)}[\log p(x|h,a,y;\theta,e) - B(x;\lambda)]$$

$$\mathrm{U}(\mathrm{x};\phi,\theta,\mathrm{w},\mathrm{e}) = E_{q(y|x;w,e)}[L(x,y;\phi,\theta,e) + \log q(y|x;w,e)]$$

第一项是不同 label 下损失的均值，第二项合起来是熵.

3. 考虑到 label 可能维度很高，通过采样的方法去估计，主要做出的改动在U这一项，下面讨论通过采样 label 后, U 的各参数求导情况，请验证:

设$\mathrm{y_i}$为根据$q(y|x;w,e)$采样一次到的 label 类别:

$$\frac{\partial}{\partial e}U(x;\phi,\theta,w,e) \approx [L(x,y_i;\phi,\theta,e)]\frac{\partial}{\partial e}(\log q(y_i|x;w,e)) + \frac{\partial}{\partial e}H(q(y|x;w,e)) + \frac{\partial}{\partial e}L(x,y_i;\phi,\theta,e)$$

$$\frac{\partial}{\partial \mathrm{w}}U(x;\phi,\theta,w,e) \approx [L(x,y_i;\phi,\theta,e)]\frac{\partial}{\partial w}(\log q(y_i|x;w,e)) + \frac{\partial}{\partial e}H(q(y|x;w,e))$$

$$\frac{\partial}{\partial(\phi,\theta)}U(x;\phi,\theta,w,e) \approx \frac{\partial}{\partial(\phi,\theta)}L(x,y_i;\phi,\theta,e)$$

$[L(x,y_i;\phi,\theta,e) - B(x;\lambda)]\frac{\partial}{\partial e}(\log q(y_i|x;w,e))$此项可以用 Policy Gradient 相关的方法处理，减去参数无关的 baseline 使得梯度更加稳定。具体函数形式？

再考虑到实际求导时候，要对一个 batch 内的所有导数取总和，因此要固定计算图中的某几个中间变量为 constant，使得梯度不反向传回，设:

$$\mathrm{CL_{i,j}} = \mathrm{L}(\mathrm{x_j},y_i;\phi,\theta,e), for\ x_j \in D_u\ y_i\ is\ sampled\ from\ q(y|x), \text{前向结果}$$

$$\mathrm{CQ_{i,j}} = \log q(y_i|x_j;w,e)\ , for\ x_j \in D_u\ y_i\ is\ sampled\ from\ q(y|x), \text{前向结果}$$

在 Theano 中，只要对 T.grad 函数中参数 consider_constant 加入以上两个计算节点，梯度就不会在这两节点上反向传播(也就是当做了常数处理)，请验证.

于是梯度仍然可以通过对一个 batch 中所有数据求导一次得到，请验证:

$$\frac{\partial}{\partial e}\sum_{\mathrm{x_j}\in D_u}U(x_j;\phi,\theta,w,e) \approx \frac{\partial}{\partial e}\left\{\sum_{x_j}\left[\left(CL_{i,j} + CQ_{i,j} + 1\right)\log q(y_i|x_j;w,e) + \mathrm{L}(\mathrm{x_j},y_i;\phi,\theta,e)\ \right]\right\}$$

其他两个参数的形式类似，注意$\sum_{\mathrm{x_j}}[\grave{}]$可以各种 tensor 操作直接得到.

$$\frac{\partial}{\partial \mathrm{w}} \sum_{\mathrm{x_j} \in D_u} U\left(x_j; \phi, \theta, w, e\right) \approx \frac{\partial}{\partial w} \left\{ \sum_{x_j} \left[\left(CL_{i,j} + CQ_{i,j} + 1 \right) \log q\left(y_i | x_j; w, e \right) \right] \right\}$$

$$\frac{\partial}{\partial(\phi, \theta)} \sum_{\mathrm{x_j} \in D_u} U\left(x_j; \phi, \theta, w, e\right) \approx \frac{\partial}{\partial(\phi, \theta)} \left\{ \sum_{x_j} \left[\mathrm{L}\left(\mathrm{x_j}, y_i; \phi, \theta, e \right) \right] \right\}$$

4. 其他部分的梯度

$$\sum_{<x,y> \in D_l} L(x, y; \phi, \theta, e) + \alpha(-\log q(y|x; w, e))$$

和先前的模型相比没有太大的变化，略.