DPM Gradient Computation

January 11, 2025

The gradient w.r.t $\mathbf{w}$ can be rewritten as

$$G_{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j\neq i}\frac{\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j)-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_j)/\tau)}{\exp(-\xi/\tau)+\sum_{j'\neq i}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_{j'})/\tau)}(\nabla_{\mathbf{w}}E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j)-\nabla_{\mathbf{w}}E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i))\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{B-1}{n-1}\sum_{j\neq i}\frac{\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j)-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_j)/\tau)}{\frac{B-1}{n-1}\exp(-\xi/\tau)+\frac{B-1}{n-1}\sum_{j'\neq i}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_{j'})/\tau)}(\nabla_{\mathbf{w}}E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j)-\nabla_{\mathbf{w}}E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)\right) \tag{1}$$

We use $s^{(i)}$ to estimate $\frac{B-1}{n-1}\sum_{j'\neq i}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_{j'})/\tau)$, which is

$$s^{(i)} \leftarrow (1-\gamma)s^{(i)}+\gamma\sum_{j\in\mathcal{B}\setminus\{i\}}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_{j'})/\tau)$$

The stochastic gradient estimator is

$$\hat{G}_{\mathbf{w}} = \frac{1}{B}\sum_{i\in\mathcal{B}}\left(\sum_{j\in\mathcal{B}\setminus\{i\}}\frac{\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j)-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_j)/\tau)}{\frac{B-1}{n-1}\exp(-\xi/\tau)+s^{(i)}}(\nabla_{\mathbf{w}}E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j)-\nabla_{\mathbf{w}}E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i))\right)$$

This corresponds to Line 592 - 633 in the code and the term $\frac{B-1}{n-1}\exp(-\xi/\tau)$ is called `offset`. The red part above is computed by `autodiff`.
The full derivative w.r.t. $\zeta^{(j)}$ can be written as

$$G_{\zeta^{(j)}} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\frac{1}{n-1}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j)-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta^{(j)})/\tau)}{\frac{1}{n-1}\sum_{j'=1}^{n}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta^{(j')})/\tau)}+\frac{1}{n}.$$

In $G_{\zeta^{(j)}}$, we need an estimator $\tilde{s}^{(i)}$ of $\frac{1}{n-1}\sum_{j'=1}^{n}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_{j'})/\tau)$. We can obtain $\tilde{s}^{(i)}$ from $s^{(i)}$, where $s^{(i)} \leftarrow (1-\gamma)s^{(i)}+\gamma\sum_{j\in\mathcal{B}\setminus\{i\}}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_{j'})/\tau)$ is the moving average estimator of $\frac{B-1}{n-1}\sum_{j'\neq i}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_{j'})/\tau)$. Note that

$$\frac{1}{n-1}\sum_{j'=1}^{n}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_{j'})/\tau)$$
$$= \frac{1}{n-1}\exp(-\zeta^{(i)}/\tau)+\frac{1}{B-1}\left(\frac{B-1}{n-1}\sum_{j'\neq i}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_{j'})-E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i)-\zeta_{j'})/\tau)\right).$$

Then, we use

$$\tilde{s}^{(i)} = \frac{1}{n-1}\exp(-\zeta^{(i)}/\tau) + \frac{1}{B-1}s^{(i)}.$$

Then, the stochastic derivative w.r.t. $\zeta^{(j)}$, $j \in \mathcal{B}$ can be defined as

$$\begin{aligned}\hat{G}_{\zeta^{(j)}} &= -\frac{1}{B}\sum_{i\in\mathcal{B}} \frac{\frac{1}{n-1}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j) - E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i) - \zeta^{(j)})/\tau)}{\tilde{s}^{(i)}} + \frac{1}{n} \\ &= -\frac{1}{B}\sum_{i\in\mathcal{B}} \frac{\frac{1}{n-1}\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j) - E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i) - \zeta^{(j)})/\tau)}{\frac{1}{n-1}\exp(-\zeta^{(i)}/\tau) + \frac{1}{B-1}s^{(i)}} + \frac{1}{n}.\end{aligned}$$

In our implementation, we rescale the stochastic derivative w.r.t. $\zeta^{(j)}$ for the convenience of tuning the learning rate. The stochastic derivative rescaled by n is

$$n\hat{G}_{\zeta^{(j)}} = -\frac{n}{n-1}\frac{1}{B}\sum_{i\in\mathcal{B}} \frac{\exp((E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_j) - E_{\mathbf{w}}(\mathbf{x}_i,\mathbf{y}_i) - \zeta^{(j)})/\tau)}{\frac{1}{n-1}\exp(-\zeta^{(i)}/\tau) + \frac{1}{B-1}s^{(i)}} + 1.$$

This corresponds to Line 645 - 675 in the code and the term $\frac{1}{n-1}\exp(-\zeta^{(i)}/\tau)$ is called `offset_tilde`.