How to Train Your Neural ODE: the World of Jacobian and Kinetic Regularization

A. Details of Section 3.1: Benamou-Brenier formulation in Lagrangian coordinates

The Benamou-Brenier formulation of the optimal transportation (OT) problem in Eulerian coordinates is

\[
\min_{\rho, f} \int_0^T \int \| f(x, t) \| \rho_t(x) \, dxdt \\
\text{subject to } \frac{\partial \rho_t}{\partial t} = - \text{div} (\rho_t f), \\
\rho_0(x) = p, \\
\rho_T(z) = q.
\]  

(18a) (18b) (18c) (18d)

The connection between continuous normalizing flows (CNF) and OT becomes transparent once we rewrite (18) in Lagrangian coordinates. Indeed, for regular enough velocity fields \( f \) one has that the solution of the continuity equation (18b), (18c) is given by \( \rho_t = z(\cdot, t)\sharp p \) where \( z \) is the flow \( z(x, t) = f(z(x, t), t), \quad z(x, 0) = x \).

The relation \( \rho_t = z(\cdot, t)\sharp p \) means that for arbitrary test function \( \phi \) we have that

\[
\int \phi(x) \rho_t(x, t) \, dx = \int \phi(z(x, t)) \, dx
\]

Therefore (18) can be rewritten as

\[
\min_f \int_0^T \int \| f(z(x, t), t) \|^2 p(x) \, dxdt \\
\text{subject to } \dot{z}(x, t) = f(z(x, t), t), \\
z(x, 0) = x, \\
z(\cdot, T)\sharp p = q.
\]  

(19a) (19b) (19c) (19d)

Note that \( \rho_t \) is eliminated in this formulation. The terminal condition (18d) is trivial to implement in Eulerian coordinates (grid-based methods) but not so simple in Lagrangian ones (19d) (grid-free methods). To enforce (19d) we introduce a penalty term in the objective function that measures the deviation of \( z(\cdot, T)\sharp p \) from \( q \). Thus, the penalized objective function is

\[
\int_0^T \int \| f(z(x, t), t) \|^2 p(x) \, dxdt + \frac{1}{\lambda} KL(z(\cdot, T)\sharp p \| q),
\]

(20)

where \( \lambda > 0 \) is the penalization strength. Next, we observe that this objective function can be written as an expectation with respect to \( x \sim p \). Indeed, the Kullback-Leibler divergence is invariant under coordinate transformations, and therefore

\[
KL(z(\cdot, T)\sharp p \| q) = KL(p \| z^{-1}(\cdot, T)\sharp q) = KL(p \| p_0) = \mathbb{E}_{x \sim p} \log \frac{p(x)}{p_0(x)} = \mathbb{E}_{x \sim p} \log p(x) - \mathbb{E}_{x \sim p} \log p_0(x)
\]

Hence, multiplying the objective function in (20) by \( \lambda \) and ignoring the \( f \)-independent term \( \mathbb{E}_{x \sim p} \log p(x) \) we obtain an equivalent objective function

\[
\mathbb{E}_{x \sim p} \left\{ \lambda \int_0^T \| f(z(x, t), t) \|^2 dt - \log p_0(x) \right\}
\]

Finally, if we assume that \( \{ x_i \}_{i=1}^N \) are iid sampled from \( p \), we obtain the empirical objective function

\[
\frac{\lambda}{N} \sum_{i=1}^N \int_0^T \| f(z(x_i, t), t) \|^2 dt - \frac{1}{N} \sum_{i=1}^N \log p_0(x_i)
\]

(22)

B. Additional results

Here we present additional generated samples on the two larger datasets considered, CelebA-HQ and ImageNet64. In addition bits/dim on clean images are reported in Table 2.

---

...
Figure 7. Quality of FFJORD RNODE generated images on ImageNet-64.

Figure 8. Quality of FFJORD RNODE generated images on CelebA-HQ. We use temperature annealing, as described in (Kingma & Dhariwal, 2018), to generate visually appealing images, with $T = 0.5, \ldots, 1$. 
Table 2. Additional results and model statistics of FFJORD RNODÉ. Here we report validation bits/dim on both validation images, and on validation images with uniform variational dequantization (ie perturbed by uniform noise). We also report number of trainable model parameters.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>BITS/DIM (CLEAN)</th>
<th>BITS/DIM (DIRTY)</th>
<th># PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>0.92</td>
<td>0.97</td>
<td>8.00e5</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>3.25</td>
<td>3.38</td>
<td>1.36e6</td>
</tr>
<tr>
<td>IMAGE.Net64</td>
<td>3.72</td>
<td>3.83</td>
<td>2.00e6</td>
</tr>
<tr>
<td>CELEBA-HQ256</td>
<td>0.72</td>
<td>1.04</td>
<td>4.61e6</td>
</tr>
</tbody>
</table>