Useful Facts About Latent-Variable Generative Models

Karl Stratos

In this note, p_X always refers to the marginal density of a latent-variable generative model p_Z and $p_{X|Z}$ where $Z \in \mathcal{Z}$.[1] So "learning p_X" means learning p_Z and $p_{X|Z}$.

1 Three Types of Estimation

We are interested in estimating the population density $\mathbf{pop}_X$ with the marginal p_X by minimizing some divergence $D(\mathbf{pop}_X, p_X)$. Three common types of the estimation problem are

$$\min_{p_X} \; D(\mathbf{pop}_X, p_X) \qquad \text{(direct marginalization)}$$
$$\min_{p_X} \; \min_{q} \; U(\mathbf{pop}_X, p_X, q) \qquad \text{(variational optimization)}$$
$$\min_{p_X} \; \max_{q} \; L(\mathbf{pop}_X, p_X, q) \qquad \text{(variational adversarial optimization)}$$

Direct marginalization. Sometimes it is possible to directly calculate the marginal. It is trivial if $\mathcal{Z}$ is small. If $\mathcal{Z}$ is a set of discrete structures such as sequences or trees, it is possible with conditional independence assumptions [3, 11]. In this case it is natural to minimize the KL divergence. Using $D_{\mathrm{KL}}(\mathbf{pop}_X || p_X) = H(\mathbf{pop}_X, p_X) - H(\mathbf{pop}_X)$

$$\min_{p_X} \; D_{\mathrm{KL}}(\mathbf{pop}_X, p_X) \equiv \min_{p_X} \; H(\mathbf{pop}_X, p_X) = \max_{p_X} \; \mathop{\mathbf{E}}_{x \sim \mathbf{pop}_X} [\log p_X(x)]$$

Variational optimization. Generally direct marginalization is intractable. In this case we can consider estimating an equivalent objective by minimizing an upper bound $U(\mathbf{pop}_X, p_X, q)$ with a variational model q such that (1) it is easy to compute, and (2) it is tight for an optimal q (henceforth tightable). For instance, EM minimizes the KL divergence by minimizing the minimization of a tightable upper bound $-\mathrm{ELBO}(\mathbf{pop}_X, p_X, q_{Z|X})$ on $H(\mathbf{pop}_X, p_X)$ where $q_{Z|X}$ estimates the intractable posterior $p_{Z|X}$ [6, 12]:

$$\min_{p_X} \; D_{\mathrm{KL}}(\mathbf{pop}_X, p_X) \equiv \min_{p_X} \; \left(\min_{q_{Z|X}} \; -\mathrm{ELBO}(\mathbf{pop}_X, p_X, q_{Z|X}) \right)$$

The consistency of p_X follows from the fact that the bound is tightable (assuming universality). The fact that the objective remains a minimization is convenient in practice.

Variational adversarial optimization. There are cases we maximize a lower bound $L(\mathbf{pop}_X, p_X, q)$ where q enjoys similar properties for estimating other divergence measures. For instance, GAN minimizes the Jensen-Shannon divergence by minimizing the maximization of a tightable upper bound $\mathrm{NCE}(\mathbf{pop}_X, p_X, q_{A|X})$ on $2\mathrm{JSD}(\mathbf{pop}_X || p_X) - \log 4$ where $q_{A|X}$ is tasked with discriminating between $\mathbf{pop}_X$ and p_X [7]:

$$\min_{p_X} \; \mathrm{JSD}(\mathbf{pop}_X || p_X) \equiv \min_{p_X} \; \left(\max_{q_{A|X}} \; \mathrm{NCE}(\mathbf{pop}_X, p_X, q_{A|X}) \right)$$

The divergence measure has been generalized to f-divergences [15] and the Wasserstein metric [1]. The consistency of p_X again follows from the fact that the bound is tightable (assuming universality). However, the objective is adversarial and more difficult to optimize in practice.

[1] We write "density" or "distribution" interchangeably to denote a probability function and $\sum$ to denote marginalization whether the considered variable is discrete or continuous.

2 Forms of ELBO

The evidence lower bound (ELBO) emerges through an effort to replace the posterior $p_{Z|X}$ with a variational model $q_{Z|X}$ in the expected log-likelihood:[2]

$$\mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X}[\log p_X(x)] = \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ z\sim q_{Z|X}(\cdot|x)}}\left[\log \frac{p_{XZ}(x,z)}{p_{Z|X}(z|x)}\frac{q_{Z|X}(z|x)}{q_{Z|X}(z|x)}\right] = \underbrace{\mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ z\sim q_{Z|X}(\cdot|x)}}\left[\log \frac{p_{XZ}(x,z)}{q_{Z|X}(z|x)}\right]}_{\mathrm{ELBO}(\mathbf{pop}_X, p_X, q_{Z|X})} + \underbrace{D_{\mathrm{KL}}(q_{Z|X}||p_{Z|X})}_{\geq 0}$$

One can write ELBO in various forms by manipulating terms:

$$\begin{aligned}
\mathrm{ELBO}(\mathbf{pop}_X, p_X, q_{Z|X}) &= \mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X}[\log p_X(x)] - D_{\mathrm{KL}}(q_{Z|X}||p_{Z|X}) && \text{(EM)} \\
&= \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ z\sim q_{Z|X}(\cdot|x)}}[\log p_{XZ}(x,z)] + H(q_{Z|X}) \\
&= \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ z\sim q_{Z|X}(\cdot|x)}}\left[\log p_{X|Z}(x|z)\right] - D_{\mathrm{KL}}(q_{Z|X}||p_Z) && \text{(VAE)} \\
&= \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ z\sim q_{Z|X}(\cdot|x)}}\left[\log p_{X|Z}(x|z)\right] - D_{\mathrm{KL}}(q_{Z|X}||q_Z) - D_{\mathrm{KL}}(q_Z||p_Z) && \text{(DVAE)} \\
&= \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ z\sim q_{Z|X}(\cdot|x)}}\left[\log p_{X|Z}(x|z)\right] + H(q_{Z|X}) - H(q_Z, p_Z) && \text{(MAXENT)}
\end{aligned}$$

The first two forms yield the traditional alternating optimization steps in the EM algorithm. The VAE form is the standard VAE objective in which the reconstruction term is estimated by sampling and the KL term is estimated in closed form. The DVAE form is the VAE form further decomposed; one can easily check that $D_{\mathrm{KL}}(q_{Z|X}||p_Z) = D_{\mathrm{KL}}(q_{Z|X}||q_Z) + D_{\mathrm{KL}}(q_Z||p_Z)$ where $q_Z(z) = \mathbf{E}_{x\sim\mathbf{pop}_X}[q_{Z|X}(z|x)]$. MAXENT is easily derived from DVAE. DVAE can be used for the following additional interpretations of VAE.

Rate-distortion autoencoders. First note that $D_{\mathrm{KL}}(q_{Z|X}||q_Z) = I(X,Z;\mathbf{pop}^q_{XZ})$ is the mutual information between X and Z under the joint density $\mathbf{pop}^q_{XZ}(x,z) = \mathbf{pop}_X(x)q_{Z|X}(z|x)$. We can view DVAE as a nested maximization over p_Z, $p_{X|Z}$, and $q_{Z|X}$. The optimal prior is always $p_Z = q_Z$ which eliminates the last KL term. The resulting optimization problem is equivalent to

$$\min_{p_{X|Z},\, q_{Z|X}} \; I(X,Z;\mathbf{pop}^q_{XZ}) + H(\mathbf{pop}^q_{X|Z}, p_{X|Z})$$

This shows the standard rate-distortion tradeoff where we want to limit the channel capacity of the encoder $q_{Z|X}$ for light-weight communication while limiting distortion $H(\mathbf{pop}^q_{X|Z}, p_{X|Z}) \leq H_{\max}$.

Disentanglement. Assume $Z = (Z_1 \ldots Z_m)$ and the model prior $p_Z = \prod_i p_{Z_i}$ is component-wise independent (usually the case). Let q_{Z_i} denote the marginal of q_Z for the i-th variable; note that q_Z may still be a complicated joint density. It is easy to verify that

$$D_{\mathrm{KL}}(q_Z||p_Z) = D_{\mathrm{KL}}\left(q_Z \middle|\middle| \prod_{i=1}^m q_{Z_i}\right) + \sum_{i=1}^m D_{\mathrm{KL}}(q_{Z_i}||p_{Z_i})$$

where the first term is also known as total correlation (a multivariate generalization of mutual information). Thus minimizing the KL term in VAE involves minimizing dependencies between latent components under q_Z. Implicit disentanglement observed in VAE (e.g., Z_{gender} vs Z_{mustache}) is attributed to this term. Several weighting schemes have been proposed to control the level of disentanglement [8, 5].

[2] One related lower bound on the log-likelihood is

$$\mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X \\ z\sim p_Z}}\left[\log p_{X|Z}(x|z)\right] \leq \mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X}\left[\log \mathop{\mathbf{E}}_{z\sim p_Z}\left[p_{X|Z}(x|z)\right]\right] = \mathop{\mathbf{E}}_{x\sim\mathbf{pop}_X}[\log p_X(x)]$$

which follows from Jensen's inequality. This objective is useful if we want to sample z from the model's own prior during learning; it can be optimized with REINFORCE in that case (Section 4) [19].

2.1 Direct Estimation of the Log-Likelihood

ELBO is just a lower bound on the log-likelihood (LL) which is what we really care about. The gap between ELBO and LL is the KL divergence between $q_{Z|X}$ and $p_{Z|X}$, which means when they are not equal a Monte Carlo estimate of ELBO (assuming fixed data x for simplicity) $(1/K)\log(p_{XZ}(x,z^{(k)})/q_{Z|X}(z^{(k)}|x))$ where $z^{(1)}\dots z^{(K)} \sim q_{Z|X}(\cdot|x)$ may still be far smaller than $\log p_X(x)$ even if $K \to \infty$. However, we can simply compute a Monte Carlo estimate of $p_X(x)$ directly, in particular using importance sampling with $q_{Z|X}$ as the proposal distribution. That is,

$$\log p_X(x) = \log \mathop{\mathbf{E}}_{z\sim q_{Z|X}(\cdot|x)}\left[\frac{p_{XZ}(x,z)}{q_{Z|X}(z|x)}\right] \approx \log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p_{XZ}(x,z^{(k)})}{q_{Z|X}(z^{(k)}|x)}\right)$$

Clearly as $K \to \infty$ the estimate converges to LL assuming bounded $\mathbf{E}_{z\sim q_{Z|X}(\cdot|x)}[p_{XZ}(x,z)/q_{Z|X}(z|x)]$. The corresponding population-level objective defined for each value of K is actually a lower bound on LL by Jensen's inequality:

$$\mathcal{L}_K = \mathop{\mathbf{E}}_{z^{(1)}\dots z^{(K)}\sim q_{Z|X}(\cdot|x)}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p_{XZ}(x,z^{(k)})}{q_{Z|X}(z^{(k)}|x)}\right)\right] \leq \log p_X(x)$$

Note that $\mathcal{L}_1$ coincides with ELBO: thus $\mathcal{L}_K$ can be viewed as a multi-sample lower bound that, unlike ELBO, converges to LL as $K \to \infty$. $\mathcal{L}_K$ can be used as an alternative training objective for learning p_{XZ} and $q_{Z|X}$ (importance weighted autoencoders (IWAs) [4]) or as a way to estimate true LL value (instead of ELBO) after learning a VAE for evaluation purposes.

3 Prior

The choice of prior can be important from an optimization perspective as well as a modeling perspective. Recall that the VAE objective is, written explicitly as a function of p_Z and $p_{X|Z}$,

$$\mathrm{ELBO}(\mathbf{pop}_X, p_Z, p_{X|Z}, q_{Z|X}) = \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X\\ z\sim q_{Z|X}(\cdot|x)}}\left[\log p_{X|Z}(x|z)\right] - D_{\mathrm{KL}}(q_{Z|X}||p_Z)$$

p_Z affects the decoder $q_{Z|X}$ through the regularization term, which in turn affects the reconstruction term. A richer prior such as multimodal instead of unimodal can help achieve better objective value.

Mixture prior. One common approach to enriching the prior is to make it a mixture distribution by introducing an additional latent variable $C \in \{1\dots K\}$ and define $p_Z(z) = \sum_{c=1}^{K} p_{Z|C}(z|c)p_C(c)$. We assume $X \perp\!\!\!\perp C|Z$ so that the model defines the joint density $p_{XZC}(x,z,c) = p_{X|Z}(x|z)p_{Z|C}(z|c)p_C(c)$. If we further define the variational posterior $q_{ZC|X}(z,c|x) = q_{Z|X}(z|x)q_{C|X}(c|x)$ with the assumption $Z \perp\!\!\!\perp C|X$, the first term in ELBO does not change and only the regularization term changes to

$$\begin{aligned} D_{\mathrm{KL}}(q_{ZC|X}||p_{ZC}) &= D_{\mathrm{KL}}(q_{C|X}||p_C) + D_{\mathrm{KL}}(q_{Z|X}||p_{Z|C}) \\ &= \mathop{\mathbf{E}}_{\substack{x\sim\mathbf{pop}_X\\ c\sim q_{C|X}(\cdot|x)}}\left[\log\frac{q_{C|X}(c|x)}{p_C(c)} + D_{\mathrm{KL}}(q_{Z|X}(\cdot|x)||p_{Z|C}(\cdot|c))\right] \end{aligned}$$

which is easy to calculate assuming small K and a closed-form solution for the KL term over Z as usual. It can be viewed as more fine-grained regularization in which we make $q_{Z|X} \approx p_{Z|C}$ where C is with respect to $q_{C|X} \approx p_C$.

Mixture prior from the variational posterior. Choosing p_Z to be a mixture distribution can be justified in terms of the optimal prior. Recall from the DVAE form of ELBO that the optimal prior is given by $\mathbf{E}_{x\sim\mathbf{pop}_X}[q_{Z|X}(z|x)]$ for any fixed $q_{Z|X}$. Thus the optimal prior for an empirical estimate of ELBO based on iid samples $x_1\dots x_N \sim \mathbf{pop}_X$ is actually the mixture distribution $(1/N)\sum_{i=1}^{N} q_{Z|X}(z|x_i)$. The mixture prior $p_Z(z) = \sum_{c=1}^{K} p_{Z|C}(z|c)p_C(c)$ can be seen as approximating this optimal prior which becomes exact when p_C is uniform over $K = N$ components and $p_{Z|C}(z|c) = q_{Z|X}(z|x_c)$. We can consider a more direct approximation by explicitly using $q_{Z|X}$ to define p_Z, for instance $p_Z(z) = (1/K)\sum_{c=1}^{K} q_{Z|X}(z|\tilde{x}_c)$ where $\tilde{x}_1\dots\tilde{x}_K$ are either random samples or learnable parameters that represent "pseudo-inputs". This parameter sharing between the prior and the decoder is shown to be potentially useful [16].

Hierarchical prior. There may be cases in which there is a natural structure to the latent variable Z. For instance, if X represents a sentence, we may think of a topic Z_1, think of facts about the topic Z_2, and then generate X. In this case we can model the joint density as $p_{XZ_1Z_2}(x, z_1, z_2) = p_{X|Z_1Z_2}(x|z_1, z_2)p_{Z_2|Z_1}(z_2|z_1)p_{Z_1}(z_1)$ and define the variational posterior as $q_{Z_1Z_2|X}(z_1, z_2|x) = q_{Z_2|XZ_1}(z_2|x, z_1)q_{Z_1|X}(z_1|x)$. This is almost the same as having a mixture prior except that we do not make conditional independence assumptions. ELBO can still be optimized with a suitable parameterization, for instance Z_1 and Z_2 are isotropic Gaussians.

Compartmentalized prior. There may also be cases in which we want to compartmentalize $Z = (Z_1, Z_2)$ with $Z_1 \perp\!\!\!\perp Z_2$. For instance, Kingma et al. (2014) define $Z_1 \in \mathbb{R}^d$ as the style and $Z_2 \in \{1 \ldots L\}$ as the label of an MNIST digit image X and consider the model $p_{XZ_1Z_2}(x, z_1, z_2) = p_{X|Z_1Z_2}(x|z_1, z_2)p_{Z_2}(z_2)p_{Z_1}(z_1)$ where the decoder $p_{X|Z_1Z_2}$ is a continuous-discrete hybrid [13]. The variational posterior $q_{Z_1Z_2|X}(z_1, z_2|x) = q_{Z_2|X}(z_2|x)q_{Z_1|XZ_2}(z_1|x, z_2)$ provides a label classifier $q_{Z_2|X}$. The model can be trained in a semi-supervised manner by jointly optimizing the ELBO of $\log p_X(x)$ on unlabeled images and $\log p_{XZ_2}(x, z_2)$ on labeled images.

Conditional prior. Assume a joint pensity $\mathbf{pop}_{XY}$ over X and Y as the population density. Consider the model $p_{XYZ}(x, y, z) = p_{X|Z}(x|z)p_{Z|Y}(z|y)\mathbf{pop}_Y(y)$ representing the generative story $Y \to Z \to X$ in which X is conditionally independent of Y given the bottleneck variable Z. We minimize $D_{\mathrm{KL}}(\mathbf{pop}_{XY}||p_{XY}(x, y))$ over the model parameters. Equivalently we maximize the conditional log likelihood $\mathbf{E}_{(x,y)\sim\mathbf{pop}_{XY}}[\log p_{X|Y}(x|y)]$. We do this by introducing a variational posterior $q_{Z|XY}$ and maximizing the ELBO lower bound

$$\max_{p_{Z|Y},\, p_{X|Z},\, q_{Z|XY}} \quad \mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathbf{pop}_{XY} \\ z\sim q_{Z|XY}(\cdot|x,y)}} \left[\log p_{X|Z}(x|z)\right] - D_{\mathrm{KL}}(q_{Z|XY}||p_{Z|Y})$$

4 Types of Non-Differentiability

Deterministic operation. Consider the step function $\textsc{step} : \mathbb{R} \to \{0, 1\}$ which outputs 1 iff the input value is nonnegative. Since it is non-differentiable at 0 and has derivative 0 almost everywhere, it is meaningless to talk about a gradient. One heuristic to obtain a meaningful gradient signal is to linearize $\textsc{step}(a) \approx a$ (which preserves the sign) in the backward pass so that

$$\frac{\partial J(\textsc{step}(f(\theta)))}{\partial\theta} = \frac{\partial J(\textsc{step}(f(\theta)))}{\partial\textsc{step}(f(\theta))}\frac{\partial\textsc{step}(f(\theta))}{\partial f(\theta)}\frac{\partial f(\theta)}{\partial\theta} \approx \frac{\partial J(\textsc{step}(f(\theta)))}{\partial\textsc{step}(f(\theta))}\frac{\partial f(\theta)}{\partial\theta}$$

We can also consider a "multidimensional step function". Define $\textsc{snap} : \mathbb{R}^K \to \{e_1 \ldots e_K\}$ by $\textsc{snap}(u) = e_{k^*}$ where $k^* = \arg\max_{k=1}^K u_k$ and $e_1 \ldots e_K \in \{0, 1\}^K$ are standard basis elements. It is non-differentiable along $(K-1)$-dimensional manifolds and has gradient 0^K almost everywhere, but we can linearize $\textsc{snap}(u) \approx u$ (which preserves the argmax) in the backward pass so that

$$\frac{\partial J(\textsc{snap}(f(\theta)))}{\partial\theta} \approx \frac{\partial J(\textsc{snap}(f(\theta)))}{\partial\textsc{snap}(f(\theta))}\frac{\partial f(\theta)}{\partial\theta} \tag{1}$$

Stochastic operation. Consider a stochastic function which outputs a certain value with a certain probability. It is only meaningful to talk about the differentiability of such a function with respect to its expectation. Let p_Z^θ denote a differentiable function of θ that defines a density over some variable Z. Given an objective function $J(z)$[3], we can consider unbiased gradient estimators such as

$$\begin{aligned}
\frac{\partial}{\partial\theta}\mathop{\mathbf{E}}_{z\sim p_Z^\theta}[J(z)] &= \sum_{z\in\mathcal{Z}} J(z)\frac{\partial}{\partial\theta}p_Z^\theta(z) && \text{(direct marginalization)} \\
&= \mathop{\mathbf{E}}_{z\sim p_Z^\theta}\left[J(z)\frac{\partial}{\partial\theta}\log p_Z^\theta(z)\right] && \text{(score function estimator)} \\
&= \mathop{\mathbf{E}}_{\epsilon\sim p_{\mathcal{E}}}\left[\frac{\partial}{\partial\theta}J(\pi^\theta(\epsilon))\right] && \text{(reparameterization trick)}
\end{aligned}$$

Direct marginalization is an option if computationally possible (e.g., $\mathcal{Z}$ is a small discrete set) [13]. The other two gradient estimates are based on sampling. The score function estimator (aka. REINFORCE [18]) is high-variance

[3] In general $J(\theta, z)$ can depend on θ through other connections.

and normally requires additional techniques to reduce variance (control variates). The reparameterization trick is an option if $z \sim p_Z^\theta$ is distributed as $z = \pi^\theta(\epsilon)$ where $\pi^\theta(\epsilon)$ is a differentiable function of θ and $\epsilon \sim p_\mathcal{E}$ is some random variable that does not depend on θ. An isotropic Gaussian density is a classical example: $z \sim \mathcal{N}(\mu_\theta, \mathrm{diag}\left(\sigma_\theta^2\right) I_d)$ is distributed as $z = \mu_\theta + \sigma_\theta \odot \epsilon$ where $\epsilon \sim \mathcal{N}(0, I_d)$ [12].

4.1 Backpropagation Through Discrete Sampling

When Z is discrete and neither direct marginalization nor the score function estimator is a good option (due to computational costs or high variance), we can consider biased gradient estimators.

Bernoulli variable. Let $Z \in \{0,1\}$ with $p_Z^\theta(1) = f(\theta)$. Combining the reparameterization trick $z = \textsc{step}(f(\theta) - \epsilon)$ where $\epsilon \sim \mathcal{U}(0,1)$ and the linear approximation $\textsc{step}(a) \approx a$ in the backward pass, we derive the straight-through gradient estimator [9, 2]:

$$\frac{\partial}{\partial\theta} \mathop{\mathbf{E}}_{z \sim p_Z^\theta}\left[J(z)\right] = \mathop{\mathbf{E}}_{\epsilon\sim p_\mathcal{E}}\left[\frac{\partial}{\partial\theta} J(\textsc{step}(f(\theta)-\epsilon))\right] \approx \mathop{\mathbf{E}}_{\epsilon\sim p_\mathcal{E}}\left[\frac{\partial J(\textsc{step}(f(\theta)-\epsilon))}{\partial \textsc{step}(f(\theta)-\epsilon)} \frac{\partial f(\theta)}{\partial\theta}\right] = \mathop{\mathbf{E}}_{z\sim p_Z^\theta}\left[\frac{\partial J(z)}{\partial z}\frac{\partial f(\theta)}{\partial \theta}\right]$$

Of course in this case direct marginalization is trivial, but this is applicable for $Z \in \{0,1\}^d$ where p_Z^θ is a product distribution. In that case direct marginalization has complexity $O(2^d)$ whereas sampling has complexity $O(d)$.

Categorical variable: Gumbel-Softmax. Let $Z \in \{1 \dots K\}$ where $K > 2$. Without loss of generality, let $Z \in \{e_1 \dots e_K\}$ be represented as a K-dimensional standard basis element with $p_Z^\theta = f(\theta) \in \Delta^{K-1}$. It can be shown that $z \sim p_Z^\theta$ (which is a vertex in Δ^{K-1}) is distributed as $z = \mathrm{softmax}((\log f(\theta) + \epsilon)/\tau)$ where $\epsilon \sim \mathrm{Gumbel}^K(0,1)$ and $\tau > 0$ as τ goes to zero, yielding the Gumbel-Softmax (GS) estimator [10, 14]

$$\frac{\partial}{\partial\theta} \mathop{\mathbf{E}}_{z\sim p_Z^\theta}\left[J(z)\right] = \lim_{\tau\to 0} \mathop{\mathbf{E}}_{\epsilon\sim\mathrm{Gumbel}^K(0,1)}\left[\frac{\partial}{\partial\theta} J\left(\mathrm{softmax}\left(\frac{\log f(\theta)+\epsilon}{\tau}\right)\right)\right]$$

In practice τ is fixed (e.g., 0.9) or annealed, so GS is biased. GS involves a K-dimensional sample and does not seem to offer any computational advantage over direct marginalization over K values. But consider:

1. Suppose the objective $J(z, z')$ involves nested sampling $z \sim \mathrm{Cat}(f(\theta))$ and $z' \sim \mathcal{N}(\mu(z), \mathrm{diag}\left(\sigma^2(z)\right) I_d)$. Direct marginalization requires drawing K conditional samples of Z' and thus $O(Kd)$ time, whereas GS gives

$$\lim_{\tau\to 0} \mathop{\mathbf{E}}_{\epsilon\sim\mathrm{Gumbel}^K(0,1)}\left[\mathop{\mathbf{E}}_{\epsilon'\sim\mathcal{N}(0,I_d)}\left[\frac{\partial}{\partial\theta} J\left(z_\theta^{\epsilon,\tau}, \mu(z_\theta^{\epsilon,\tau}) + \sigma(z_\theta^{\epsilon,\tau}) \odot \epsilon'\right)\right]\right]$$

   where $z_\theta^{\epsilon,\tau} = \mathrm{softmax}((\log f(\theta)+\epsilon)/\tau) \in \Delta^{K-1}$. This estimator requires one conditional sample of Z' and thus $O(d)$ time.

2. Suppose $Z \in \{1 \dots K\}^d$ and p_Z^θ is a product distribution over d dimensions. Then direct marginalization has complexity $O(K^d)$ whereas sampling has complexity $O(Kd)$.

We can consider a straight-through version of GS for cases in which we need a discrete sample by snapping in the forward pass and softmaxing in the backward pass:

$$\frac{\partial}{\partial\theta} \mathop{\mathbf{E}}_{\epsilon\sim\mathrm{Gumbel}^K(0,1)}\left[J\left(\textsc{snap}\left(\frac{\log f(\theta)+\epsilon}{\tau}\right)\right)\right] \approx \mathop{\mathbf{E}}_{\epsilon\sim\mathrm{Gumbel}^K(0,1)}\left[\frac{\partial}{\partial\theta} J\left(\mathrm{softmax}\left(\frac{\log f(\theta)+\epsilon}{\tau}\right)\right)\right]$$

Categorical variable: vector quantization Let $f(\theta) \in \mathbb{R}^d$ and assume $C \in \mathbb{R}^{K\times d}$. We discretize $f(\theta) \in \mathbb{R}^d$ into $\{1 \dots K\}$ by treating the rows of C as centroids in k-means. Estimating the gradient in this case using the straight-through estimator for snapping (1) is called vector quantization (VQ):

$$\frac{\partial J(\textsc{snap}(u(\theta, C))}{\partial\theta} \approx \frac{\partial J(\textsc{snap}(u(\theta, C)))}{\partial \textsc{snap}(u(\theta, C))} \frac{\partial f(\theta))}{\partial\theta}$$

where $u_i(\theta, C) = -\,||f(\theta) - C_i||$. When we want to update C as well we can add additional objectives such as minimizing $||f(\theta) - C_{i^*}||^2$ where $i^* = \arg\max_{i=1}^K u_i(\theta, C)$ [17].

References

[1] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia. PMLR.

[2] Bengio, Y., Léonard, N., and Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

[3] Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., and Klein, D. (2010). Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590. Association for Computational Linguistics.

[4] Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

[5] Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620.

[6] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[8] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, **2**(5), 6.

[9] Hinton, G. (2012). Neural networks for machine learning. coursera,[video lectures].

[10] Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

[11] Kim, Y., Dyer, C., and Rush, A. M. (2019). Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385.

[12] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

[13] Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589.

[14] Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

[15] Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279.

[16] Tomczak, J. and Welling, M. (2018). Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223.

[17] van den Oord, A., Vinyals, O., *et al.* (2017). Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315.

[18] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, **8**(3-4), 229–256.

[19] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.