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 $Z \in \mathcal{Z}$. 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 $\text{pop}_X$ with the marginal $p_X$ by minimizing some divergence $D(\text{pop}_X, p_X)$. Three common types of the estimation problem are

\[
\begin{align*}
\min_{p_X} D(\text{pop}_X, p_X) & \quad \text{(direct marginalization)} \\
\min_{p_X} \min_q U(\text{pop}_X, p_X, q) & \quad \text{(variational optimization)} \\
\min_{p_X} \max_q L(\text{pop}_X, p_X, q) & \quad \text{(variational adversarial optimization)}
\end{align*}
\]

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 $\text{KL}(\text{pop}_X || p_X) = H(\text{pop}_X, p_X) - H(\text{pop}_X)$

\[
\min_{p_X} \text{KL}(\text{pop}_X, p_X) \equiv \min_{p_X} H(\text{pop}_X, p_X) = \max_{p_X} \mathbb{E}_{x \sim \text{pop}_X} [\log p_X(x)]
\]

Variational optimization. Generally direct marginalization is intractable. In this case we can consider estimating an equivalent objective by minimizing an upper bound $U(\text{pop}_X, p_X, q)$ 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 maximization of a tightable upper bound $-\text{ELBO}(\text{pop}_X, p_X, q_{Z|X})$ on $H(\text{pop}_X, p_X)$ where $q_{Z|X}$ estimates the intractable posterior $p_{Z|X}$ [6, 12]:

\[
\min_{p_X} \text{KL}(\text{pop}_X, p_X) \equiv \min_{p_X} \left(\min_{q_{Z|X}} -\text{ELBO}(\text{pop}_X, p_X, q_{Z|X})\right)
\]

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

\[
\min_{p_X} \text{JSD}(\text{pop}_X || p_X) \equiv \min_{p_X} \left(\max_{q_{A|X}} \text{NCE}(\text{pop}_X, p_X, q_{A|X})\right)
\]

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:\footnote{One related lower bound on the log-likelihood is}

$$\mathbb{E}_{x \sim \text{pop}_X} \left[ \log p_X(x) \right] = \mathbb{E}_{x \sim \text{pop}_X} \left[ \log \frac{p_{XZ}(x, z) q_{Z|X}(z|x)}{p_{Z|X}(z|x) q_{Z|X}(z|x)} \right] = \mathbb{E}_{x \sim \text{pop}_X, z \sim q_{Z|X}(z|x)} \left[ \log p_{XZ}(x, z) q_{Z|X}(z|x) \right] + D_{KL}(q_{Z|X} || p_{Z|X}) \geq 0$$

One can write ELBO in various forms by manipulating terms:

$$\text{ELBO}(\text{pop}_X, p_X, q_{Z|X}) = \mathbb{E}_{x \sim \text{pop}_X} \left[ \log p_X(x) \right] - D_{KL}(q_{Z|X} || p_{Z|X})$$

(EM)

$$= \mathbb{E}_{x \sim \text{pop}_X, z \sim q_{Z|X}(z|x)} \left[ \log p_X(z, x) + H(q_{Z|X}) \right]$$

(EM)

$= \mathbb{E}_{x \sim \text{pop}_X, z \sim q_{Z|X}(z|x)} \left[ \log p_{X|Z}(z|x) - D_{KL}(q_{Z|X} || p_{Z}) \right]$ (VAE)

$= \mathbb{E}_{x \sim \text{pop}_X, z \sim q_{Z|X}(z|x)} \left[ \log p_{X|Z}(z|x) \right] - D_{KL}(q_{Z|X} || q_{Z}) - D_{KL}(q_{Z} || p_{Z})$ (DVAE)

$= \mathbb{E}_{x \sim \text{pop}_X, z \sim q_{Z|X}(z|x)} \left[ \log p_{X|Z}(z|x) \right] + H(q_{Z|X}) - H(q_{Z}, p_{Z})$ (MAXENT)

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 $D_{KL}(q_{Z|X} || p_Z) = D_{KL}(q_{Z|X} || q_Z) + D_{KL}(q_Z || p_Z)$ where $q_{Z}(z) = \mathbb{E}_{x \sim \text{pop}_X} [q_{Z|X}(z|x)].$ MAXENT is easily derived from DVAE. DVAE can be used for the following additional interpretations of VAE.

Rate-distortion autoencoders. First note that $D_{KL}(q_{Z|X} || q_Z) = I(X, Z; \text{pop}_X^q(Z))$ is the mutual information between $X$ and $Z$ under the joint density $\text{pop}_X^q(x, z) = \text{pop}_X(x)q_{Z|X}(z|x).$ 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

$$\min_{p_{X|Z}, p_Z, q_{Z|X}} I(X, Z; \text{pop}_X^q(Z)) + H(\text{pop}_X^q(Z), p_X|Z)$$

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 $H(\text{pop}_X^q(Z), p_X|Z) \leq H_{\text{max}}.$

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

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

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., $Z_{\text{gender}}$ vs $Z_{\text{mustache}}$) is attributed to this term. Several weighting schemes have been proposed to control the level of disentanglement [8, 5].
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 \( z^{(1)} \ldots z^{(K)} \sim q_{Z|X}(\cdot|x) \) 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,

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

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

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

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 \( p_{X|Z} \),

\[
\text{ELBO}(\text{pop}_X, p_Z, p_{X|Z}, q_{Z|X}) = \mathbb{E}_{z \sim \text{pop}_X} \left[ \log p_{X|Z}(x|z) \right] - D_{KL}(q_{Z|X}||p_Z)
\]

\( p_Z \) 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 \( C \in \{1 \ldots K\} \) and define \( p_C(z) = \sum_{c=1}^{K} p_{Z|C}(z|c)p_C(c) \). We assume \( X \perp \!\!\!\!\perp C|Z \) so that the model defines the joint density \( p_{XZC}(x,z,c) = p_X(x|z)p_{Z|C}(z|c)p_C(c) \). If we further define the variational posterior \( q_{C|X}(z,c|x) = q_{Z|X}(z|x)q_{C|X}(c|x) \) with the assumption \( Z \perp \!\!\!\!\perp C|X \), the first term in ELBO does not change and only the regularization term changes to

\[
D_{KL}(q_{Z|C|X}||p_ZC) = D_{KL}(q_{C|X}||p_C) + D_{KL}(q_{Z|X}(\cdot|c)||p_{Z|C}(\cdot|c))
\]

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 \( \mathbb{E}_{z \sim \text{pop}_X} [q_{Z|X}(z|x)] \) for any fixed \( q_{Z|X} \). Thus the optimal prior for an empirical estimate of ELBO based on iid samples \( x_1 \ldots x_N \sim \text{pop}_X \) 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 \( K \) 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 \( p_Z(z) = (1/K) \sum_{c=1}^{K} q_{Z|X}(z|x_c) \) where \( \tilde{x}_1 \ldots \tilde{x}_K \) 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 construct the joint density as \( p_{XZ_1Z_2}(x, z_1, z_2) = p_X(z_1, z_2)p_{Z_1}(z_1)p_{Z_2|Z_1}(z_2|z_1) \) and define the variational posterior as \( q_{Z_1Z_2}(z_1, z_2|x) = q_{Z_1}(z_1|x)q_{Z_2|Z_1}(z_2|z_1) \). 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 \( p_{XZ_1Z_2}(x, z_1, z_2) = p_X(z_1, z_2)p_{Z_1|Z_2}(z_1|z_2)p_{Z_2}(z_2) \) where the decoder \( p_{X|Z_1Z_2} \) is a continuous-discrete hybrid [13]. The variational posterior \( q_{Z_1Z_2}(z_1, z_2|x) = q_{Z_2}(z_2|x)q_{Z_1|Z_2}(z_1|z_2) \) provides a label classifier \( q_{Z_1|x} \). 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 density \( p_{XY|Z}(x, y, z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)p_{Z}(z) \) representing the generative story \( Y \to Z \to X \) in which \( X \) is conditionally independent of \( Y \) given the bottleneck variable \( Z \). We minimize \( D_{KL}(p_{XY} || p_{XY|Z}) \) over the model parameters. Equivalently, we maximize the conditional log likelihood \( \mathbb{E}_{(x, y, z) \sim p_{XY}} \left[ \log p_{X|Z}(x|z) \right] - D_{KL}(q_{Z|XY} || p_{Z|Y}) \). We do this by introducing a variational posterior \( q_{Z|XY} \) and maximizing the ELBO lower bound

$$
\max_{p_{Z|Y}, p_{X|Z}, q_{Z|XY}} \mathbb{E}_{(x, y, z) \sim p_{XY}(x, y)} \left[ \log p_{X|Z}(x|z) \right] - D_{KL}(q_{Z|XY} || p_{Z|Y})
$$

4 Types of Non-Differentiability

Deterministic operation. Consider the step function \( \text{step} : \mathbb{R} \to \{0, 1\} \) 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 \( \text{step}(a) \approx a \) (which preserves the sign) in the backward pass so that

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

We can also consider a “multidimensional step function”. Define \( \text{snap} : \mathbb{R}^K \to \{e_1 \ldots e_K\} \) by \( \text{snap}(u) = e_k^* \) 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 almost everywhere, but we can linearize \( \text{snap}(u) \approx u \) (which preserves the argmax) in the backward pass so that

$$
\frac{\partial J(\text{snap}(f(\theta)))}{\partial \theta} \approx \frac{\partial J(\text{snap}(f(\theta)))}{\partial \text{snap}(f(\theta))} \cdot \frac{\partial f(\theta)}{\partial \theta}
$$

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 \( p_\theta^Z \) denote a differentiable function of \( \theta \) that defines a density over some variable \( Z \). Given an objective function \( J(z)^3 \), we can consider unbiased gradient estimators such as

$$
\frac{\partial}{\partial \theta} \mathbb{E}_{z \sim p_\theta^Z} [J(z)] = \sum_{z \in Z} J(z) \frac{\partial}{\partial \theta} p_\theta^Z(z) \quad \text{(direct marginalization)}
$$

$$
= \mathbb{E}_{z \sim p_\theta^Z} \left[ J(z) \frac{\partial}{\partial \theta} \log p_\theta^Z(z) \right] \quad \text{(score function estimator)}
$$

$$
= \mathbb{E}_{\epsilon \sim \mathcal{U}} \left[ \frac{\partial}{\partial \theta} J(\pi^\theta(\epsilon)) \right] \quad \text{(reparameterization trick)}
$$

Direct marginalization is an option if computationally possible (e.g., \( Z \) 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 \( \theta \) through other connections.
and normally requires additional techniques to reduce variance (control variates). The reparameterization trick is an option if \( z \sim p^\theta_Z \) is distributed as \( z = \pi^\theta(\epsilon) \) where \( \pi^\theta(\epsilon) \) is a differentiable function of \( \theta \) and \( \epsilon \sim p^\epsilon \) is some random variable that does not depend on \( \theta \). An isotropic Gaussian density is a classical example: \( z \sim \mathcal{N}(\mu_\theta, \text{diag}(\sigma^2_\theta)) I_d \) 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 \( p^\theta_Z(1) = f(\theta) \). Combining the reparameterization trick \( z = \text{STEP}(f(\theta) - \epsilon) \) and the linear approximation \( \text{STEP}(a) \approx a \) in the backward pass, we derive the straight-through gradient estimator [9, 2]:

\[
\frac{\partial}{\partial \theta} \mathbb{E}_{z \sim p^\epsilon_Z} [J(z)] = \mathbb{E}_{\epsilon \sim p^\epsilon} \left[ \frac{\partial}{\partial \theta} \text{STEP}(f(\theta) - \epsilon) \right] \approx \mathbb{E}_{\epsilon \sim p^\epsilon} \left[ \frac{\partial J(\text{STEP}(f(\theta) - \epsilon))}{\partial \text{STEP}(f(\theta) - \epsilon)} \frac{\partial f(\theta)}{\partial \theta} \right] = \mathbb{E}_{z \sim p^\epsilon_Z} \left[ \frac{\partial J(z)}{\partial z} \frac{\partial f(\theta)}{\partial \theta} \right]
\]

Of course in this case direct marginalization is trivial, but this is applicable for \( Z \in \{0, 1\}^d \) where \( p^\theta_Z \) 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 \( Z \in \{1 \ldots K\} \) where \( K > 2 \). Without loss of generality, let \( Z \in \{e_1 \ldots e_K\} \) be represented as a \( K \)-dimensional standard basis element with \( p^\theta_Z = f(\theta) \in \Delta^{K-1} \). It can be shown that \( z \sim p^\theta_Z \) (which is a vertex in \( \Delta^{K-1} \)) is distributed as \( z = \text{softmax}((\log f(\theta) + \epsilon)/\tau) \) where \( \epsilon \sim \text{Gumbel}(0, 1) \) and \( \tau > 0 \) as \( \tau \) goes to zero (Appendix A), yielding the Gumbel-Softmax (GS) estimator [10, 14]

\[
\frac{\partial}{\partial \theta} \mathbb{E}_{z \sim p^\epsilon_Z} [J(z)] = \lim_{\tau \to 0} \mathbb{E}_{\epsilon \sim \text{Gumbel}(0, 1)} \left[ \frac{\partial}{\partial \theta} J \left( \text{softmax} \left( \frac{\log f(\theta) + \epsilon}{\tau} \right) \right) \right]
\]

In practice \( \tau \) 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 \( z \sim \text{Cat}(f(\theta)) \) and \( z' \sim \mathcal{N}(\mu(z), \text{diag}(\sigma^2(z)) I_d) \). Direct marginalization requires drawing \( K \) conditional samples of \( Z' \) and thus \( O(Kd) \) time, whereas GS gives

\[
\lim_{\tau \to 0} \mathbb{E}_{\epsilon \sim \text{Gumbel}(0, 1)} \left[ \mathbb{E}_{\epsilon' \sim \mathcal{N}(0, I_d)} \left[ \frac{\partial}{\partial \theta} J \left( z^\epsilon_{\theta}, \mu(z^\epsilon_{\theta}) + \sigma(z^\epsilon_{\theta}) \odot \epsilon' \right) \right] \right]
\]

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

2. Suppose \( Z \in \{1 \ldots K\}^d \) and \( p^\theta_Z \) is a product distribution over \( d \) dimensions. Then direct marginalization has complexity \( O(Kd) \) 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 softmating in the backward pass:

\[
\frac{\partial}{\partial \theta} \mathbb{E}_{\epsilon \sim \text{Gumbel}(0, 1)} \left[ J \left( \text{SNAP} \left( \frac{\log f(\theta) + \epsilon}{\tau} \right) \right) \right] \approx \mathbb{E}_{\epsilon \sim \text{Gumbel}(0, 1)} \left[ \frac{\partial}{\partial \theta} J \left( \text{softmax} \left( \frac{\log f(\theta) + \epsilon}{\tau} \right) \right) \right]
\]

**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 \( \{1 \ldots K\} \) 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):

\[
\frac{\partial J(\text{SNAP}(u(\theta, C)))}{\partial \theta} \approx \frac{\partial J(\text{SNAP}(u(\theta, C)))}{\partial \text{SNAP}(u(\theta, C))} \frac{\partial f(\theta)}{\partial \theta}
\]

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


A The Gumbel Distribution

The Gumbel distribution with location \( \mu \in \mathbb{R} \) and scale \( \beta > 0 \) has the CDF over \( \mathbb{R} \):

\[
\text{Gumbel}(\mu, \beta)(x) = \int_{-\infty}^{x} \text{Gumbel}(\mu, \beta)(t)dt = \exp\left(-\exp\left(-\frac{x - \mu}{\beta}\right)\right)
\]

where \( \mu = 0 \) and \( \beta = 1 \) yields the standard form. Differentiating the CDF gives the PDF:

\[
\text{Gumbel}(\mu, \beta)(x) = \frac{1}{\beta} \exp\left(-\frac{x - \mu}{\beta} - \exp\left(-\frac{x - \mu}{\beta}\right)\right)
\]

which is admittedly complicated with nested exponentials.\(^5\) Gumbel is closed under linear transformation: if \( Z \sim \text{Gumbel}(0, 1) \), then \( \mu + \beta Z \sim \text{Gumbel}(\mu, \beta) \).\(^6\) If \( z \sim \text{Gumbel}(\mu, \beta) \) and \( z' \sim \text{Gumbel}(\mu', \beta) \), then \( z - z' \sim \text{Logistic}(\mu - \mu', \beta) \). The Gumbel distribution with location \( \mu \) and scale \( \beta > 0 \) has the CDF over \( \mathbb{R} \):

\[
\text{Gumbel}(\mu, \beta)(x) = \exp\left(-\exp\left(-\frac{x - \mu}{\beta}\right)\right)
\]

A The Gumbel Distribution

\[
\text{Gumbel}(\mu, \beta)(x) = \int_{-\infty}^{x} \text{Gumbel}(\mu, \beta)(t)dt = \exp\left(-\exp\left(-\frac{x - \mu}{\beta}\right)\right)
\]

where \( \mu = 0 \) and \( \beta = 1 \) yields the standard form. Differentiating the CDF gives the PDF:

\[
\text{Gumbel}(\mu, \beta)(x) = \frac{1}{\beta} \exp\left(-\frac{x - \mu}{\beta} - \exp\left(-\frac{x - \mu}{\beta}\right)\right)
\]

which is admittedly complicated with nested exponentials.\(^5\) Gumbel is closed under linear transformation: if \( Z \sim \text{Gumbel}(0, 1) \), then \( \mu + \beta Z \sim \text{Gumbel}(\mu, \beta) \).\(^6\) If \( z \sim \text{Gumbel}(\mu, \beta) \) and \( z' \sim \text{Gumbel}(\mu', \beta) \), then \( z - z' \sim \text{Logistic}(\mu - \mu', \beta) \). The Gumbel distribution with location \( \mu \) and scale \( \beta > 0 \) has the CDF over \( \mathbb{R} \):

\[
\text{Gumbel}(\mu, \beta)(x) = \exp\left(-\exp\left(-\frac{x - \mu}{\beta}\right)\right)
\]

A.1 The Gumbel-Max Trick

Assume an integer \( K \geq 1 \) and logits \( u \in \mathbb{R}^K \). Let \( X \sim \text{Cat} (\text{softmax}(u)) \). Let \( \epsilon_1 \ldots \epsilon_K \sim \text{Gumbel}(0, 1) \) and define \( Y := \arg\max_{k=1}^K u_k + \epsilon_k \).\(^8\)

Theorem A.1 (The Gumbel-max trick). \( \Pr(X = k) = \Pr(Y = k) \) for all \( k = 1 \ldots K \).

The proof is trivial for \( K = 1 \) and simple for \( K = 2 \).\(^9\) The proof for general \( K \) is less simple.

\(^4\)If \( F(x) = \int_{-\infty}^{x} f(t)dt \) is a CDF of a PDF \( f \) with support on \( x \geq c \), and \( G \) is any antiderivative of \( f \) (i.e., \( G'(x) = f(x) \)), the fundamental theorem of calculus says \( F(x) = G(x) - G(c) \). Therefore, \( F'(x) = \frac{\partial}{\partial x} (G(x) - G(c)) = G'(x) = f(x) \).

\(^5\)It can be verified that the mean is \( \mu + \gamma \beta \) where \( \gamma = 0.5772 \ldots \) is the Euler-Mascheroni constant, the variance is \( \frac{\pi^2}{6} \beta^2 \).

\(^6\)The CDF of \( \mu + \beta Z \) is \( \Pr(\mu + \beta Z \leq x) = \Pr(Z \leq \frac{x - \mu}{\beta}) = \exp\left(-\exp\left(-\frac{x - \mu}{\beta}\right)\right) \). But this is exactly the CDF of \( \text{Gumbel}(\mu, \beta) \).

\(^7\)Recall \( \text{Exp}_\lambda(x) = [x \geq 0] \lambda \exp(-\lambda x) \) is the continuous version of the geometric distribution that represents how long we have to wait until an event with rate \( \lambda \) happens. It has the CDF: \( 1 - \exp(-\lambda x) \).

\(^8\)\( u_k + \epsilon_k \in \mathbb{R} \) are distinct with probability 1 since \( \epsilon_1 \ldots \epsilon_K \) are drawn iid from a continuous distribution.

\(^9\)WLOG let \( k = 1 \). Using the fact that the CDF of \( \text{Logistic}(0, 1) \) is the sigmoid function, we have

\[
\Pr(Y = 1) = \int_{-\infty}^{\infty} \text{Pr}(u \sim \text{Gumbel}(0, 1), u_1 + \epsilon_1 \geq u_2 + \epsilon_2) = \int_{-\infty}^{\infty} \text{Pr}(z \sim \text{Logistic}(0, 1), z \leq u_1 - u_2) = \sigma(u_1 - u_2)
\]

By the usual relation between the sigmoid and the softmax when \( K = 2 \), we have

\[
\sigma(u_1 - u_2) = \frac{1}{1 + \exp(u_2 - u_1)} = \frac{\exp(u_1)}{\exp(u_1) + \exp(u_2)} = \Pr(X = 1)
\]
Proof. WLOG let \( k = 1 \).

\[
\Pr(Y = 1) = \Pr_{\epsilon_1 \ldots \epsilon_K \sim \text{Gumbel}(0,1)}(u_1 + \epsilon_1 \geq u_k + \epsilon_k \ \forall k > 1)
\]

\[
= \mathbf{E}_{\epsilon_1 \sim \text{Gumbel}(0,1)} \left[ \prod_{k > 1} \Pr_{\epsilon_k \sim \text{Gumbel}(0,1)}(u_1 + \epsilon_1 \geq u_k + \epsilon_k) \right]
\]

\[
= \mathbf{E}_{\epsilon_1 \sim \text{Gumbel}(0,1)} \left[ \prod_{k > 1} \Pr_{\epsilon_k \sim \text{Gumbel}(0,1)}(\epsilon_k \leq u_1 - u_k + \epsilon_1) \right]
\]

\[
= \mathbf{E}_{\epsilon_1 \sim \text{Gumbel}(0,1)} \left[ \prod_{k > 1} \exp(-\exp(-u_1 + u_k - \epsilon_1)) \right]
\]

\[
= \int_{-\infty}^{\infty} \exp(-\epsilon_1 - \exp(-\epsilon_1)) \prod_{k > 1} \exp(-\exp(-u_1 + u_k - \epsilon_1)) d\epsilon_1
\]

\[
= \int_{-\infty}^{\infty} \exp\left(-\epsilon_1 - \left(\sum_{k=1}^{K} \exp(-u_1 + u_k)\right) \exp(-\epsilon_1)\right) d\epsilon_1
\]

By \( https://www.integral-calculator.com \), for \( C > 0 \),

\[
\int_{-\infty}^{\infty} \exp(-x - C \exp(-x)) \, dx = \frac{\exp(-C \exp(-x))}{C} \bigg|_{-\infty}^{\infty} = \frac{1}{C}
\]

Thus the above integral evaluates to

\[
\frac{1}{\sum_{k=1}^{K} \exp(-u_1 + u_k)} = \frac{\exp(u_1)}{\sum_{k=1}^{K} \exp(u_k)} = \Pr(X = 1)
\]

\( \square \)

**Gumbel-Softmax.** We observe that for any \( \epsilon \in \mathbb{R}^K \) the distribution

\[
\delta_{\tau} := \text{softmax}\left(\frac{u + \epsilon}{\tau}\right) \in [0, 1]^K
\]

converges to the one-hot vector representation of \( \kappa^* = \arg\max_{k=1}^{K} u_k + \epsilon_k \) as \( \tau \to 0^+ \). Thus if \( \epsilon_1 \ldots \epsilon_K \sim \text{Gumbel}(0,1) \), then \( \delta_{\tau} \) is distributed as the one-hot vector representation of \( X \sim \text{Cat} (\text{softmax}(u)) \) as \( \tau \to 0^+ \) by Theorem A.1.

**Lemma A.2.** For any \( c > 0 \),

\[
\lim_{N \to \infty} \left(1 - \frac{1}{N}c\right)^N = \lim_{\epsilon \to 0^-} (1 + \epsilon c)^{-\frac{1}{\epsilon}} = e^{-c}
\]

Proof. The first equality follows by the change of variable \( N = -\frac{1}{\epsilon} \) and taking the one-sided limit \( \epsilon \to 0^- \). Taking negative log on both sides of the second equality gives

\[
-\log \lim_{\epsilon \to 0^-} (1 + \epsilon c)^{-\frac{1}{\epsilon}} = \lim_{\epsilon \to 0^-} \frac{\log (1 + \epsilon c)}{\epsilon} = c
\]

where the first equality holds since log is continuous at \( (1 + \epsilon c)^{-\frac{1}{\epsilon}} > 0 \) for all \( \epsilon_{\min} < \epsilon < 0 \) for some \( \epsilon_{\min} \). Define \( f(x) = \log(1 + xc) \) and note that

\[
f'(x) := \lim_{\epsilon \to 0} \frac{f(x + \epsilon) - f(x)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\log(1 + (x + \epsilon)c) - \log(1 + xc)}{\epsilon} \Rightarrow f'(0) = \lim_{\epsilon \to 0} \frac{\log(1 + \epsilon c)}{\epsilon}
\]

\( f(x) \) is uniformly continuous at \( x = 0 \) so the one-sided limit is the same as the two-sided limit. Hence the claim (2) is equivalent to \( f'(0) = c \). This follows since \( f'(x) = c/(1 + xc) \) by the chain rule on log. \( \square \)