Control Variate

Richard Xu

December 4, 2020

1 Control variate motivations

variance reduction works by modifying "function of a random variable" so that its expectation remains same, but variance reduces

in this artcile, our primary interest is in estimating a derivative:

$$\nabla_\theta \mathbb{E}_{p(b|\theta)}[f(b)] \tag{1}$$

1.1 control variate C

We illustrate this through Reinforcement Learning, letting $b \to \tau$, $p \to \pi$ and $f(b) \to R(\tau)$:

1.1.1 Reinforcement Learning

$$\begin{aligned}
J(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta(\tau)}\big[R(\tau)\big] \\
\implies \nabla_\theta J(\theta) &= \nabla_\theta \int_\tau \pi_\theta(\tau) R(\tau) = \int_\tau \nabla_\theta \pi_\theta(\tau) R(\tau) \\
&= \int_\tau \nabla_\theta \big[\log(\pi_\theta(\tau))\big] \pi_\theta(\tau) R(\tau) \\
&= \mathbb{E}_{\tau \sim \pi_{\theta(\tau)}}\big[\nabla_\theta \log(\pi_\theta(\tau))\big] R(\tau)
\end{aligned}$$

now adding a control variate C :

$$\begin{aligned}
&= \mathbb{E}_{\tau \sim \pi_{\theta(\tau)}}\big[\nabla_\theta \log(\pi_\theta(\tau))\big(R(\tau) - C\big)\big] \\
&= \mathbb{E}_{\tau \sim \pi_{\theta(\tau)}}\big[\nabla_\theta \log(\pi_\theta(\tau)) R(\tau)\big] - \underbrace{\mathbb{E}_{\tau \sim \pi_{\theta(\tau)}}\big[\nabla_\theta \log(\pi_\theta(\tau)) C\big]}_{=0}
\end{aligned} \tag{2}$$

so what doe the RHS of Eq. (2) $= 0$?

$$\begin{aligned}
&\mathbb{E}_{\tau \sim \pi_{\theta(\tau)}}\big[\nabla_\theta \log(\pi_\theta(\tau)) C\big] \\
&= \int_\tau \nabla_\theta \log(\pi_\theta(\tau)) C \times \pi_{\theta(\tau)} \\
&= C \int_\tau \underbrace{\nabla_\theta \big[\log(\pi_\theta(\tau))\big] \pi_\theta(\tau)}_{\nabla_\theta \pi_\theta(\tau)} \\
&= C \nabla_\theta \int_\tau \pi_\theta(\tau) = 0
\end{aligned} \tag{3}$$

This means $\times C$ won't matter. This means that a constant C, i.e., **independent** of both θ and b will generate an unbiased control variate.

1.1.2 when C is a function of θ

however, when C is independent of τ, but dependent of θ:

$$
\begin{aligned}
\nabla_\theta \int_\tau C(\theta)\pi(\tau|\theta)\mathrm{d}\tau &= \nabla_\theta C(\theta) \int_\tau \pi(\tau|\theta)\mathrm{d}\tau \\
&= C'(\theta)
\end{aligned} \tag{4}
$$

2 More sophisticated Control Variate

We can add a more generic, biased control variate C. Since we know that when C is a function of b, then $\mathbb{E}_{p(b|\theta)}\big[C\ \nabla_\theta \log p(b|\theta)\big]$ is unbiased. Therefore, the biased C must be a function of b. Since it's biased, we need to delete it:

$$
\begin{aligned}
&\nabla_\theta \mathbb{E}_{p(b|\theta)}\big[f(b)\big] \\
&= \nabla_\theta \mathbb{E}_{p(b|\theta)}\big[f(b) - C + C\big] \\
&= \nabla_\theta \mathbb{E}_{p(b|\theta)}\big[f(b) - C\big] + \nabla_\theta \mathbb{E}_{p(b|\theta)}\big[C\big] \\
&= \mathbb{E}_{p(b|\theta)}\Big[(f(b) - C)\nabla_\theta \log(p(b|\theta))\Big] + \nabla_\theta \mathbb{E}_{p(b|\theta)}\big[C\big] \\
&= \mathbb{E}_{p(b|\theta)}\Big[f(b)\nabla_\theta \log p(b|\theta)\Big] - \mathbb{E}_{p(b|\theta)}\Big[C\ \nabla_\theta \log p(b|\theta)\Big] + \underbrace{\nabla_\theta \mathbb{E}_{p(b|\theta)}\big[C\big]}_{\mathbb{E}_{p(b|\theta)}\left[C\ \nabla_\theta \log p(b|\theta)\right]} \\
&= \mathbb{E}_{p(b|\theta)}\Big[f(b)\nabla_\theta \log p(b|\theta)\Big] - \mathbb{E}_{p(b|\theta)}\Big[C\ \nabla_\theta \log p(b|\theta)\Big] + \mathbb{E}_{p(b|\theta)}\big[C\ \nabla_\theta \log p(b|\theta)\big]
\end{aligned} \tag{5}
$$

substitute $C \equiv C_\phi(z)$:

$$
\begin{aligned}
\nabla_\theta \mathbb{E}_{p(b|\theta)}\big[f(b)\big] &= \mathbb{E}_{p(b|\theta)}\Big[f(b)\nabla_\theta \log p(b|\theta)\Big] - \mathbb{E}_{p(b|\theta)}\Big[C_\phi(z)\ \nabla_\theta \log p(b|\theta)\Big] + \nabla_\theta \mathbb{E}_{p(b|\theta)}\big[C_\phi(z)\big] \\
&= \underbrace{\mathbb{E}_{p(b|\theta)}\Big[f(b)\nabla_\theta \log p(b|\theta)\Big]}_{\nabla_\theta \hat{f}_{\text{Reinforce}}} - \underbrace{\mathbb{E}_{p(b|\theta)}\Big[C_\phi(z)\ \nabla_\theta \log p(b|\theta)\Big]}_{\nabla_\theta \hat{C}_{\text{Reinforce}}\ \text{without re-parameterization}} + \underbrace{\nabla_\theta \mathbb{E}_{p(z)}\big[C_\phi(z)\big]}_{\nabla_\theta \hat{C}_{\text{Reinforce}}\ \text{with re-parameterization}}
\end{aligned} \tag{6}
$$

3 More sophisticated control variate: more the correlation, the better!

(Paisley ICML12) introduce a control variate $g(x)$ approximates $f(x)$ well

when closed-form $\mathbb{E}[g(\theta)]$ isn't possible, a low variance approximate is used:

$$
\hat{f}(x) = f(x) - h\Big(\underbrace{g(x) - \mathbb{E}[g(x)]}_{\mathbb{E}\left[g(x) - \mathbb{E}[g(x)]\right] = 0}\Big) \tag{7}
$$

knowing:

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j:\ j>i}^{n} a_i a_j \text{Cov}(X_i, X_j) \tag{8}$$

try it on $\text{Var}(\hat{f})$:

$$\begin{aligned} \text{Var}(\hat{f}) &= \text{Var}(f) - 2h\ \text{Cov}(f,g) + h^2\text{Var}(g) \\ \implies \nabla_h \text{Var}(\hat{f}) &= -2\ \text{Cov}(f,g) + 2h\ \text{Var}(g) = 0 \\ \implies h^* &= \frac{\text{Cov}(f,g)}{\text{Var}(g)} \end{aligned} \tag{9}$$

substitute h^*:

$$\begin{aligned} \text{Var}(\hat{f}) &= \text{Var}(f) - 2\frac{\text{Cov}(f,g)}{\text{Var}(g)}\ \text{Cov}(f,g) + \frac{\text{Cov}(f,g)^2}{\text{Var}(g)^2}\text{Var}(g) \\ &= \text{Var}(f) - 2\frac{\text{Cov}(f,g)^2}{\text{Var}(g)} + \frac{\text{Cov}(f,g)^2}{\text{Var}(g)} \\ &= \text{Var}(f) - \frac{\text{Cov}(f,g)^2}{\text{Var}(g)} \\ \implies \frac{\text{Var}(\hat{f})}{\text{Var}(f)} &= 1 - \frac{\text{Cov}(f,g)^2}{\text{Var}(g)^2} = 1 - \text{Corr}(f,g) \end{aligned} \tag{10}$$

meaning, higher the correlation between f and g, the less the variance of $\hat{f}$

4 RELAX algorithm:

the control variate we have added is:

$$C \equiv \mathbb{E}_{p(\tilde{z}|b,\theta)}[c_\phi(\tilde{z})] \tag{11}$$

note that C is a function of b, through $p(\tilde{z}|b,\theta)$:

$$\begin{aligned} \nabla_\theta \mathbb{E}_{p(b|\theta)}[f(b)] = \mathbb{E}_{p(b|\theta)}\Big[f(b)\nabla_\theta \log p(b|\theta)\Big] - \mathbb{E}_{p(b|\theta)}\Big[\mathbb{E}_{p(\tilde{z}|b,\theta)}[c_\phi(\tilde{z})]\nabla_\theta \log p(b|\theta)\Big] \\ - \mathbb{E}_{p(b|\theta)}\Big[\nabla_\theta \mathbb{E}_{p(\tilde{z}|b,\theta)}\big[c_\phi(\tilde{z})\big]\Big] + \nabla_\theta \mathbb{E}_{p(z|\theta)}\big[c_\phi(z)\big] \end{aligned} \tag{12}$$

we can equally write $\mathbb{E}_{p(\tilde{z}|b,\theta)}[c_\phi(\tilde{z})]$ as $\mathbb{E}_{p(z|b,\theta)}[c_\phi(z)]$, however, $\tilde{z}$ emphasis that it is sampled from posterior $p(\tilde{z}|b)$condition on b, whereas z is from prior $p(z)$

the posterior term $p(\tilde{z}|b,\theta)$ highlights that $\tilde{z}$ and b are correlated! The same does **not** seen in prior $p(z)$

note that RELAX estimator does not mean it correspond to a particular expression for C in Eq.(6)

4.1 Why the above is unbiased

imagine we can prove:

$$\mathbb{E}_{p(b|\theta)}\left[\mathbb{E}_{p(\tilde{z}|b,\theta)}[c_\phi(\tilde{z})]\nabla_\theta \log p(b|\theta)\right] = -\mathbb{E}_{p(b|\theta)}\left[\nabla_\theta \mathbb{E}_{p(\tilde{z}|b,\theta)}\left[c_\phi(\tilde{z})\right]\right] + \nabla_\theta \mathbb{E}_{p(z|\theta)}\left[c_\phi(z)\right] \tag{13}$$

or,

$$\nabla_\theta \mathbb{E}_{p(z|\theta)}\left[c_\phi(z)\right] = \mathbb{E}_{p(b|\theta)}\left[\nabla_\theta \mathbb{E}_{p(\tilde{z}|b,\theta)}\left[c_\phi(\tilde{z})\right]\right] + \mathbb{E}_{p(b|\theta)}\left[\mathbb{E}_{p(\tilde{z}|b,\theta)}[c_\phi(\tilde{z})]\nabla_\theta \log p(b|\theta)\right] \tag{14}$$

then Eq.(25) is unbiased estimator and we are all done.

To show this, we first need under **Deterministic function:** it has a very special property when $p(b|z)$ is deterministic:

$$\begin{aligned} p(z) &\equiv p(z|b)p(b) \quad \text{or} \\ p(z|\theta) &\equiv p(z|b,\theta)p(b|\theta) \end{aligned} \tag{15}$$

with that

$$\begin{aligned} \nabla_\theta \mathbb{E}_{p(z|\theta)}\left[c_\phi(z)\right] &= \mathbb{E}_{p(z|\theta)}\left[\nabla_\theta \log(p(z|\theta))c_\phi(z)\right] \\ &= \mathbb{E}_{p(b|\theta)}\left[\mathbb{E}_{p(\tilde{z}|b,\theta)}\left[c_\phi(z)\nabla_\theta \log(p(z|\theta))\right]\right] \quad \text{using Eq.(15)} \\ &= \mathbb{E}_{p(b|\theta)}\left[\mathbb{E}_{p(\tilde{z}|b,\theta)}\left[c_\phi(z)\nabla_\theta \log\left(p(z|\theta) + p(z|b,\theta)\right)\right]\right] \quad \text{using Eq.(15) again!} \\ &= \mathbb{E}_{p(b|\theta)}\left[\mathbb{E}_{p(\tilde{z}|b,\theta)}\left[c_\phi(z)\nabla_\theta \log p(z|\theta)\right]\right] + \mathbb{E}_{p(b|\theta)}\left[\mathbb{E}_{p(\tilde{z}|b,\theta)}\left[c_\phi(z)\nabla_\theta \log p(z|b,\theta)\right]\right] \\ &= \mathbb{E}_{p(b|\theta)}\left[\nabla_\theta \mathbb{E}_{p(\tilde{z}|b,\theta)}\left[c_\phi(\tilde{z})\right]\right] + \mathbb{E}_{p(b|\theta)}\left[\mathbb{E}_{p(\tilde{z}|b,\theta)}[c_\phi(\tilde{z})]\nabla_\theta \log p(b|\theta)\right] \end{aligned} \tag{16}$$

also, for $\mathbb{E}_{p(b|\theta)}\left[f(b)\nabla_\theta \log p(b|\theta)\right]$:

$$
\begin{aligned}
\mathbb{E}_{p(z)}\Big[f(H(z))\nabla_\theta \log p(z)\Big] &= \mathbb{E}_{p(b)}\Big[\mathbb{E}_{p(z|b)}\Big[f(H(z))\nabla_\theta \log p(z)\Big]\Big] \\
&= \mathbb{E}_{p(b)}\Big[\mathbb{E}_{p(z|b)}\Big[f(b)\nabla_\theta \log p(z)\Big]\Big] \\
&= \mathbb{E}_{p(b)}\Big[f(b)\mathbb{E}_{p(z|b)}\Big[\nabla_\theta \log p(z)\Big]\Big] \\
&= \mathbb{E}_{p(b)}\Big[f(b)\mathbb{E}_{p(z|b)}\Big[\nabla_\theta \log(p(z|b)) + \nabla_\theta \log(p(b))\Big]\Big] \\
&= \mathbb{E}_{p(b)}\Big[f(b)\underbrace{\mathbb{E}_{p(z|b)}\left[\nabla_\theta \log\big(p(z|b)\big)\right]}\Big] + \mathbb{E}_{p(b)}\left[f(b)\nabla_\theta \log(p(b))\right] \\
&= \mathbb{E}_{p(b)}\Big[f(b)\underbrace{\int_z \nabla_\theta p(z|b)}_{=0}\Big] + \nabla_\theta \mathbb{E}_{p(b)}\left[f(b)\right] \\
&= \nabla_\theta \mathbb{E}_{p(b)}\left[f(b)\right] \\
&= \mathbb{E}_{p(b|\theta)}\left[f(b)\nabla_\theta \log p(b|\theta)\right]
\end{aligned} \tag{17}
$$

Add Baseline trick to the problem

$$\begin{aligned} J(\theta) &= \mathbb{E}_{\tau\sim\pi_{\theta(\tau)}}\big[R(\tau)\big] \\ \implies \nabla_\theta J(\theta) &= \mathbb{E}_{\tau\sim\pi_{\theta(\tau)}}\big[\nabla_\theta \log \pi_\theta(\tau) R(\tau)\big] - \mathbb{E}_{\tau\sim\pi_{\theta(\tau)}}\big[C\nabla_\theta \log \pi_\theta(\tau)\big] \end{aligned} \tag{18}$$

replace $\pi_\theta(\tau)$ with $p(b|\theta)$ and $r(\tau)$ with $f(b)$

$$\nabla_\theta J(\theta) = \mathbb{E}_{b\sim p(b|\theta)}\big[\nabla_\theta \log p(b|\theta) f(b)\big] - \mathbb{E}_{b\sim p(b|\theta)}\big[C\nabla_\theta \log p(b|\theta)\big] \tag{19}$$

what should a good expression of C be?
Variance Reduction through control variate
apply this sophisticated control variate:

$$\begin{aligned} \nabla_\theta \mathbb{E}_{p(b,z)}[f(b)] &= \nabla_\theta \Big(\mathbb{E}_{p(b,z)}[f(b) - g(z)] + \mathbb{E}_{p(b,z)}[g(z)]\Big) \\ &= \mathbb{E}_{p(b,z)}\Big[\big(f(b) - g(z)\big)\nabla_\theta \log p(b)\Big] + \nabla_\theta \mathbb{E}_{p(b,z)}[g(z)] \\ \nabla_\theta \mathbb{E}_{p(b)}[f(b)] &= \mathbb{E}_{p(b)}\Big[f(b)\nabla_\theta \log p(b)\Big] - \mathbb{E}_{p(b,z)}\Big[g(z)\nabla_\theta \log p(b)\Big] + \nabla_\theta \mathbb{E}_{p(z)}[g(z)] \end{aligned} \tag{20}$$

a good choice of $g(z)$ is important, remember corr(f, g) needs to be high

$$g(z) \equiv \mathbb{E}_{p(z|b)}\big[f(\sigma_\lambda(z))\big] \tag{21}$$

after some simplification:

$$\begin{aligned} &\nabla_\theta \mathbb{E}_{p(b)}[f(b)] \\ &= \mathbb{E}_{p(b)}\Big[(f(b)\nabla_\theta \log p(b)\Big] - \mathbb{E}_{p(b)}\Big[\mathbb{E}_{p(z|b)}\big[f(\sigma_\lambda(z))\big]\nabla_\theta \log p(b)\Big] + \nabla_\theta \mathbb{E}_{\underbrace{p(z|b)}_{p(z)}}\big[f(\sigma_\lambda(z))\big] \end{aligned} \tag{22}$$

change of variables
this is:

$$p(z) = \left|\frac{d\epsilon}{dz}\right| p(\epsilon) \implies |p(z)dz| = |p(\epsilon)d\epsilon| \tag{23}$$

$$\begin{aligned} \nabla_\theta \mathbb{E}_{p(z;\theta)}[f(z)] &= \nabla_\theta \int p(z;\theta) f(z) dz \\ = \nabla_\theta \int p(\epsilon) f(z) d\epsilon &= \nabla_\theta \int p(\epsilon) f(g(\epsilon,\theta)) d\epsilon \\ = \nabla_\theta \mathbb{E}_{p(\epsilon)}[f(g(\epsilon,\theta))] &= \mathbb{E}_{p(\epsilon)}[\nabla_\theta f(g(\epsilon,\theta))] \end{aligned} \tag{24}$$

5 REBAR algorithm

5.1 Put REBAR in action

we want to estimate $\nabla_\theta \mathbb{E}_{p(b|\theta)}[f(b)]$, the estimator used is:

$$
\begin{aligned}
&\nabla_\theta \mathbb{E}_{p(b|\theta)}[f(b)] \\
&= \mathbb{E}_{p(b|\theta)}\Big[f(b)\nabla_\theta \log p(b|\theta)\Big] - \mathbb{E}_{p(b|\theta)}\Big[\mathbb{E}_{p(\tilde{z}|b,\theta)}[c_\phi(\tilde{z})]\nabla_\theta \log p(b|\theta)\Big] \\
&\qquad - \mathbb{E}_{p(b|\theta)}\Big[\nabla_\theta \mathbb{E}_{p(\tilde{z}|b,\theta)}\big[c_\phi(\tilde{z})\big]\Big] + \nabla_\theta \mathbb{E}_{p(z|\theta)}\big[c_\phi(z)\big]
\end{aligned} \tag{25}
$$

remember you may substitute any function of $c_\phi(\tilde{z})$. In REBAR, one substitute:

$$
c_\phi(\tilde{z}) \equiv \sigma_\lambda(\tilde{z}) \tag{26}
$$

5.1.1 What is $\sigma_\lambda(\tilde{z}(\theta, u))$

In **un-relaxed** version, you can obtain b from $z(\theta, u)$ by:

$$
b = H(z(\theta, u)) \tag{27}
$$

In REBAR, you can obtain a **relaxed** version:

$$
\tilde{b} = \sigma_\lambda(z(\theta, u)) = \frac{1}{1 + \exp(-z/\lambda)} \tag{28}
$$

$$
\begin{aligned}
&\nabla_\theta \mathbb{E}_{p(b)}[f(b)] \\
&= \mathbb{E}_{p(b)}\Big[(f(b)\nabla_\theta \log p(b)\Big] - \mathbb{E}_{p(b)}\Big[\mathbb{E}_{p(z|b)}\big[f(\sigma_\lambda(z))\big]\nabla_\theta \log p(b)\Big] \\
&\qquad - \mathbb{E}_{p(b)}\nabla_\theta \mathbb{E}_{p(z|b)}\big[f(\sigma_\lambda(z))\big] + \mathbb{E}_{p(b)}\big[\mathbb{E}_{p(z|b)}\big[f(\sigma_\lambda(z))\nabla_\theta \log(p(b))\big]\big]
\end{aligned} \tag{29}
$$

5.2 REBAR in action

5.2.1 Under a Bernoulli setting

illustrate the case where we need to estimate $\nabla_\theta \mathbb{E}_{p(b)}[f(b)]$ where $b \sim \text{Bernoulli}(\theta)$

from detailed derivation in re-parameterization section, we see re-parameterization of Logistic random variable can be done by:

$$
\begin{aligned}
u &\sim \mathcal{U}(0, 1) \\
z(\theta, u) &= \log\left(\frac{\theta}{1-\theta}\right) + \log\left(\frac{u}{1-u}\right)
\end{aligned} \tag{30}
$$

so REBAR version is:

$$
\begin{aligned}
\tilde{b} = \sigma_\lambda(z(\theta, u)) &= \frac{1}{1 + \exp(-z/\lambda)} \\
&= \frac{1}{1 + \exp\left[-\left(\log\left(\frac{\theta}{1-\theta}\right) - \log\left(\frac{u}{1-u}\right)\right)/\lambda\right]}
\end{aligned} \tag{31}
$$

after **re-parameterization**:

$$
\begin{aligned}
&\nabla_\theta \mathbb{E}_{p(b,z)}[f(b)] \\
&= \mathbb{E}_{u\sim\mathcal{U}}\left[f(H((z))\nabla_\theta \log(p(z))\right] - \mathbb{E}_{u\sim\mathcal{U}}\left[\mathbb{E}_{v\sim\mathcal{U}}\left[f(\sigma_\lambda(\tilde{z}))\right]\nabla_\theta \log p(b)\right] \\
&\qquad - \underbrace{\mathbb{E}_{p(b)}\left[\mathbb{E}_{v\sim\mathcal{U}}\left[\nabla_\theta f(\sigma_\lambda(\tilde{z}))\right]\right]}_{\text{using } \tilde{z}} + \underbrace{\nabla_\theta \mathbb{E}_{p(b)}\left[\mathbb{E}_{v\sim\mathcal{U}}\left[f(\sigma_\lambda(z))\right]\right]}_{\text{using } z}
\end{aligned} \tag{32}
$$

5.2.2 REBAR algorithm

the algorithm is like, at each iteration, the derivative is calculated as:

$$
\begin{aligned}
&\nabla_\theta \mathbb{E}_{p(b)}\left[f(b)\right] \\
=&\mathbb{E}_{p(u,v)}\left[\left(f(b) - f(\sigma_\lambda(\tilde{z}))\right)\nabla_\theta \log p(b|\theta) - \left(\nabla_\theta f(\sigma_\lambda(\tilde{z})) - \nabla_\theta f(\sigma_{\tilde{z}})\right)\right]
\end{aligned} \tag{33}
$$

basically, $u \to z \to b$, then having $(b, v) \to \tilde{z}$

6 re-parameterization of $p(\mathbf{z}|\mathbf{b})$ in general case

we look at even more generically case of k, instead of binary case

this means knowing $\mathbf{z} = [z_1, \dots z_n]$, a vector of Gumbel random variables,

1. one-hot vector $\mathbf{b} = [0, 0, \dots, \underbrace{1}_{k^{\text{th}}}, 0, \dots]$
2. probability vector $\mathbf{p}$
3. CDF sampling variables $\mathbf{v}$

what is the linking function $\tilde{g}$ which outputs $\tilde{\mathbf{z}}$?

$$\tilde{\mathbf{z}} = \tilde{g}(\mathbf{v}, \mathbf{b}, \mathbf{p}) \tag{34}$$

obviously the z_k correspond to 1 in one-hot vector should be the largest!

6.1 Gumbel distribution properties

1. Gumbel's first wonderful property derived from its CDF:

$$\mathcal{G}\text{A}(g; \phi) = \exp(-\exp(-g + \phi))$$

then:

$$\begin{aligned}
&\implies \mathcal{G}\text{A}(g; \phi)\, \mathcal{G}\text{A}(g; \gamma) = \mathcal{G}\text{A}\Big(g; \log\big(\exp(\phi) + \exp(\gamma)\big)\Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{product of two Gumbel CDFs is a Gumbel CDF!} \\
&\implies \prod_{i=1}^{n} \mathcal{G}\text{A}(g; \phi_i) = \mathcal{G}\text{A}\Bigg(g; \log\Big(\underbrace{\sum_{i=1}^{n} \exp(\phi_i)}_{\mathcal{Z}}\Big)\Bigg) = \mathcal{G}\text{A}\big(g; \log(\mathcal{Z})\big)
\end{aligned} \tag{35}$$

in words, product of multiple Gumbel random variables are also Gumbel, with parameters are the **logSumExp** of individual parameters

2. Gumbel's second wonderful property:

$$\underbrace{\text{ga}(g; \phi)}_{\text{pdf}} = \exp(-g + \phi)\, \underbrace{\mathcal{G}\text{A}(g; \phi)}_{\text{cdf}} \tag{36}$$

6.2 Formulation for $p(\mathbf{z}|\mathbf{b})$

We know from Gumbel-Max trick that:

Given a set of Gumbel RVs:

$$\mathbf{z} = \big\{z_1 \sim \text{ga}(\log(p_1)), \dots, z_n \sim \text{ga}(\log(p_n))\big\} \tag{37}$$

probability that k^{th} element is their maximum is equal probability of sampling:

$$k \sim (p_1, \dots, p_n), \text{i.e.,} \ p_k \tag{38}$$

condition on the prior:

$$p(\mathbf{z}) = \prod_{i=1}^{n} \text{ga}(z_i; \log(p_i)) \tag{39}$$

the conditional density that the k^{th} of one-hot vector $\mathbf{b}$ is one, i.e., $b_k = 1, \ b_{i \neq k} = 0 \ \forall i$ is:

$$b_i = H_i(\mathbf{z}) \begin{cases} 1 \text{ if } z_i \geq z_j \ \forall j \neq i \\ 0 \text{ otherwise} \end{cases} \tag{40}$$

where H is the "argmax" operator

$b_k = 1$ means one-hot $\mathbf{b}$ vector having k^{th} element one, the conditional:

$$p(b_k = 1|\mathbf{z}) = \mathbf{1}(z_k \geq z_i) \tag{41}$$

computing posterior using usual:

$$\begin{aligned} p(\mathbf{z}|b_k = 1) &= \frac{p(\mathbf{z}, b_k = 1)}{p(b_k = 1)} \\ &= \frac{p(b_k = 1|\mathbf{z})p(\mathbf{z})}{p(b_k = 1)} \end{aligned} \tag{42}$$

note that this is different to computing softmax, where we need to integrate out z_k

$$
\begin{aligned}
p(b_k = 1|\mathbf{z})p(\mathbf{z}) &= \underbrace{\prod_{i=1}^{n} \mathrm{ga}(z_i; \log(p_i))}_{p(\mathbf{z})} \underbrace{\mathbf{1}(z_k \geq z_i)}_{p(b_k=1|\mathbf{z})} \\
&= \frac{\mathcal{G}\mathrm{A}(z_k; \log(1-p_k))}{\mathcal{G}\mathrm{A}(z_k; \log(1-p_k))} \prod_{i=1}^{n} \mathrm{ga}(z_i; \log(p_i)) \mathbf{1}(z_k \geq z_i) \\
&= \frac{\mathrm{ga}(z_k; \log(p_k)) \mathcal{G}\mathrm{A}(z_k; \log(1-p_k))}{\mathcal{G}\mathrm{A}(z_k; \log(1-p_k))} \prod_{i \neq k}^{n} \mathrm{ga}(z_i; \log(p_i)) \mathbf{1}(z_k \geq z_i) \\
&= \mathrm{ga}(z_k; \log(p_k)) \mathcal{G}\mathrm{A}(z_k; \log(1-p_k)) \prod_{i \neq k}^{n} \frac{\mathrm{ga}(z_i; \log(p_i)) \mathbf{1}(z_k \geq z_i)}{\mathcal{G}\mathrm{A}(z_k; \log(1-p_k))}
\end{aligned} \tag{43}
$$

using **Property 1**: Eq.(35):

$$
\begin{aligned}
\mathcal{G}\mathrm{A}\big(z_k; \log(1-p_k)\big) &= \mathcal{G}\mathrm{A}\Big(z_k; \log\Big(\sum_{i \neq k} p_i\Big)\Big) \\
&= \mathcal{G}\mathrm{A}\left(z_k; \underbrace{\log \sum_{i \neq k} \exp}_{\text{logSumExp}} \big(\underbrace{\log(p_i)}_{\phi_i}\big)\right) \\
&= \prod_{i \neq k}^{n} \mathcal{G}\mathrm{A}\big(z_k; \log(p_i)\big)
\end{aligned} \tag{44}
$$

$$
\begin{aligned}
&= \underbrace{\mathrm{ga}(z_k;\log(p_k))}\,\mathcal{G}\mathrm{A}(z_k;\log(1-p_k))\prod_{i\neq k}^{n}\frac{\mathrm{ga}(z_i;\log(p_i))\,\mathbf{1}(z_k\geq z_i)}{\mathcal{G}\mathrm{A}(z_k;\log(p_i))}\\
&= \underbrace{\exp(-z_k+\log(p_k))\mathcal{G}\mathrm{A}(z_k;\log(p_k))}_{\textbf{Property2}:\ \text{Eq.(36)}\ \mathrm{ga}(g;\phi)=\exp(-g+\phi)\mathcal{G}\mathrm{A}(g;\phi)}\ \mathcal{G}\mathrm{A}(z_k;\log(1-p_k))\prod_{i\neq k}^{n}\frac{p(z_i;\log(p_i))\mathbf{1}(z_k\geq z_i)}{\mathcal{G}\mathrm{A}(z_k;\log(p_i))}\\
&= \underline{\exp(-z_k+\log(p_k))}\,\underbrace{\mathcal{G}\mathrm{A}(z_k;\log(p_k))\mathcal{G}\mathrm{A}(z_k;\log(1-p_k))}\prod_{i\neq k}^{n}\frac{p(z_i;\log(p_i))\mathbf{1}(z_k\geq z_i)}{\mathcal{G}\mathrm{A}(z_k;\log(p_i))}\\
&= \underline{p_k\exp(-z_k)}\,\underbrace{\mathcal{G}\mathrm{A}(z_k;\log(1))}_{\textbf{Property1}:\text{Eq.(35)}}\prod_{i\neq k}^{n}\frac{p(z_i;\log(p_i))\mathbf{1}(z_k\geq z_i)}{\mathcal{G}\mathrm{A}(z_k;\log(p_i))}\\
&= p_k\,\underbrace{\exp(-z_k)\mathcal{G}\mathrm{A}(z_k;0)}\prod_{i\neq k}^{n}\frac{\mathrm{ga}(z_i;\log(p_i))\mathbf{1}(z_k\geq z_i)}{\mathcal{G}\mathrm{A}(z_k;\log(p_i))} \qquad \log(1)=0\\
&= p_k\,\underbrace{\mathrm{ga}(z_k;0)}_{\text{Eq.(36)}\ \mathrm{ga}(g;\phi)=\exp(-g+\phi)\mathcal{G}\mathrm{A}(g;\phi)}\prod_{i\neq k}^{n}\frac{\mathrm{ga}(z_i;\log(p_i))\mathbf{1}(z_k\geq z_i)}{\mathcal{G}\mathrm{A}(z_k;\log(p_i))}\\
&= \underbrace{p_k}_{p(b_k=1)}\underbrace{p_0(z_k)\prod_{i\neq k}^{n}\frac{\mathrm{ga}(z_i;\log(p_i))\mathbf{1}(z_k\geq z_i)}{\mathcal{G}\mathrm{A}(z_k;\log(p_i))}}_{p(\mathbf{z}|b_k=1)}\\
&= p(\mathbf{z}|p_k=1)\,\mathrm{Pr}(b_k=1)
\end{aligned}
\tag{45}
$$

Gumbel-max trick tells us that $\mathrm{Pr}(b_k=1)=p_k$, so we have the conditional density required for the linking function:

6.2.1 sample $p(\mathbf{z}|p_k=1)$ via re-parameterization

from uniform random variables $\mathbf{v}$, the link function to obtain $\mathbf{z}$ is:

$$
\mathbf{z}=\tilde{g}(\mathbf{v},\mathbf{b},\mathbf{p}) \tag{46}
$$

we can obtain the link function by looking at the conditional:

$$
p(\mathbf{z}|b_k=1)=\mathrm{ga}(z_k;0)\prod_{i\neq k}^{n}\frac{\mathrm{ga}(z_i;\log(p_i))\mathbf{1}(z_k\geq z_i)}{\mathcal{G}\mathrm{A}(z_k;\log(p_i))} \tag{47}
$$

link function is different for z_k and $z_{k\neq i}$:

$$
z_i=\begin{cases}\mathcal{G}\mathrm{A}^{-1}(v_i,0)\ \text{if} & i=k\\ \mathcal{G}\mathrm{A}^{-1}_{\text{Truncated}}(v_i,\log(p_i),z_k) & \text{otherwise}\end{cases} \tag{48}
$$

$$
\mathcal{G}\mathrm{A}^{-1}(z_k;v_k,0)=-\log(-\log(v_k)) \tag{49}
$$

and we know:

$$\mathcal{G}\mathrm{A}_{\text{truncate}}{}^{-1}(u, \phi, T) = \phi - \log\big(\exp(\phi - T) - \log(u)\big) \tag{50}$$

let

$$\begin{aligned} &\phi = \log(p_i) \\ &T = z_k = -\log(-\log(v_k)) \end{aligned} \tag{51}$$

then:

$$\begin{aligned}
&\mathcal{G}\mathrm{A}_{\text{truncate}}{}^{-1}(z_i; v_i, \log(p_i), z_k) \\
&= \log(p_i) - \log\Big(\exp\big(\log(p_i) + \log(-\log(v_k))\big) - \log(u)\Big) \\
&= \log(p_i) - \log\Big(\exp\Big(\log\Big(\frac{p_i}{\log(v_k)}\Big)\Big) - \log(u)\Big) \\
&= -\log\Big(\frac{1}{p_i}\Big) - \log\Big(\frac{p_i}{\log(v_k)} - \log(u)\Big) \\
&= -\log\Big(\Big(\frac{1}{p_i}\Big)\frac{p_i}{\log(v_k)} - \Big(\frac{1}{p_i}\Big)\log(u)\Big) \\
&= -\log\Big(-\log(v_k) - \frac{\log(u)}{p_i}\Big)
\end{aligned} \tag{52}$$

6.2.2 relationship with how Gumbel-max trick is caculated

key point to remember, the above require the explicit values of $z_{i\neq k}$, therefore, it uses $p(\log(p_i))$ whereas Gumbel-max trick is calculated without the explicit value of $z_{i\neq k}$, therefore it uses $\Pr(\log(p_i))$

6.2.3 Formulation for $p(\mathbf{z}|\mathbf{b})$ under recursive truncation τ

another problem is to add in further truncation τ (this is not used in this setting), in here, no requirement to obtain explicit value of $z_{i\neq k}$, so we use $\mathcal{G}\mathrm{A}(.)$

and we perform the same trick, i.e., write down joint density:

$$
\begin{aligned}
&p(\text{max element} = b, \text{max value} = z|\tau) \\
&= p\Big(G(b) = z, G(b) \geq \max_{i \neq b} G(i)|\tau\Big) = p(z)p(b|z) \\
&= \underbrace{\frac{\text{ga}(z;\phi_k)\mathbf{1}(z \leq \tau)}{\mathcal{G}\text{A}(\tau;\phi_k)}}_{\Pr\left(G(i^* \equiv b)=z|\tau\right)} \times \qquad \underbrace{\prod_{i \neq b} \frac{\mathcal{G}\text{A}(z;\phi_i)}{\mathcal{G}\text{A}(\tau;\phi_i)}}_{\Pr\left(G(b) \geq \max_{i \neq b} G(i) \middle| G(b)=z,\tau\right)} \\
&= \exp(-z+\phi_k)\mathbf{1}(z \leq \tau) \quad \times \quad \left(\frac{\mathcal{G}\text{A}(z;\phi_k)}{\mathcal{G}\text{A}(\tau;\phi_k)} \quad \times \quad \prod_{i \neq b}^{n} \frac{\mathcal{G}\text{A}(z;\phi_i)}{\mathcal{G}\text{A}(\tau;\phi_i)}\right) \quad \text{using } p_\phi(z) = \exp(-z+\phi)\mathcal{G}\text{A}(z;\phi) \\
&= \frac{\exp(\phi_k)}{\mathcal{Z}} \exp(-z + \log \mathcal{Z})\mathbf{1}(z \leq \tau) \frac{\mathcal{G}\text{A}(z;\log(\mathcal{Z}))}{\mathcal{G}\text{A}(\tau;\log(\mathcal{Z}))} \qquad \text{using } \prod_{i=1}^{n} \mathcal{G}\text{A}(z;\phi_i) = \mathcal{G}\text{A}(z;\log(\mathcal{Z})) \\
&= \frac{\exp(\phi_k)}{\mathcal{Z}} \frac{\overbrace{\exp(-z + \log \mathcal{Z})\mathcal{G}\text{A}(z;\log(\mathcal{Z}))}^{p_\phi(z)=\exp(-z+\phi)\mathcal{G}\text{A}(z;\phi)} \mathbf{1}(z \leq \tau)}{\mathcal{G}\text{A}(\tau;\log(\mathcal{Z}))} \\
&= \underbrace{\frac{\exp(\phi_k)}{\mathcal{Z}}}_{\Pr(b)} \underbrace{\frac{\text{ga}(z;\log(\mathcal{Z}))}{\mathcal{G}\text{A}(\tau;\log(\mathcal{Z}))}\mathbf{1}(z \leq \tau)}_{p(z|b,\tau)} \\
&= \Pr(b)p(z|b,\tau)
\end{aligned}
\tag{53}
$$

6.3 Re-parameterization in Bernoulli variable

just like the Gumbel-Max trick

$$
\Pr\big(k : k \sim (\text{softmax}(\Phi_1, \ldots, \Phi_k))\big) = \Pr\big(k = \max(\Phi_1 + z_1, \ldots, \Phi_k + z_k)\big) \\ \text{where } z_i \sim \text{ga}(0,1)
\tag{54}
$$

in a similar way, instead of sampling $b \sim \text{Bernoulli}(\theta)$, we can use **re-parameterization** of $u \sim \mathcal{U}(0,1)$

$$
\underbrace{\Pr\big(b = 1|\theta\big)}_{=\theta} = \Pr\left(b = H\bigg(\underbrace{\log\left(\frac{\theta}{1-\theta}\right) + \log\left(\frac{u}{1-u}\right)}_{z(\theta,u)}\bigg)\right) \\
\text{where } u \sim U(0,1) \qquad H(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}
\tag{55}
$$

in fact, the above is a specialized Gumbel-Max trick of binary cases, where H is the "max" operator

6.3.1 Re-parameterization of binary case via Logistic Distribution

Logistic Distribution has the following properties:

$$\text{PDF}: \quad p(\mu, s) = \frac{\exp^{-\frac{x-\mu}{s}}}{s\left(1+\exp^{-\frac{x-\mu}{s}}\right)^2} \tag{56}$$

$$\text{CDF}: \quad \Pr(\mu, s) = \frac{1}{1+\exp^{-\frac{x-\mu}{s}}} \tag{57}$$

$$\begin{aligned}
\text{CDF}^{-1}: \quad u &= \frac{1}{1+\exp^{-\frac{x-\mu}{s}}} \\
\Longrightarrow \frac{1}{u} &= 1+\exp^{-\frac{x-\mu}{s}} \Longrightarrow \frac{1}{u} - 1 = \exp^{-\frac{x-\mu}{s}} \\
\Longrightarrow \log\left(\frac{1}{u}-1\right) &= -\frac{x-\mu}{s} \\
\Longrightarrow x &= -\log\left(\frac{1}{u}-1\right)s+\mu = -\log\left(\frac{1-u}{u}\right)s+\mu \\
x &= \big(\log u - \log(1-u)\big)s+\mu
\end{aligned} \tag{58}$$

6.3.2 Obtain linking function $z = g(\theta, u)$ via Logistic Distribution

property: if $z_1 \sim \text{ga}(\mu_1, \beta)$ and $z_2 \sim \text{ga}(\mu_2, \beta)$, then:

$$z = z_1 - z_2 \sim \text{Logistic}(\mu_1 - \mu_2, \beta) \tag{59}$$

difference of two Gumbel RVs is a Logistic RV!

in binary case, i.e., $K = 2$, max of K Gumbels, becomes "max of two Gumbel random variables" with locations $\mu_1 = \log \alpha_1$ and $\mu_2 = \log \alpha_2$ respectively, and also let $\beta = 1$

$$U \sim \mathcal{U}(0,1), \quad \text{then} \qquad z = z_1 - z_2 = \log U - \log(1-U) + \log \alpha_1 - \log \alpha_2 \tag{60}$$

$$b = \begin{cases} 1 & \text{if} \quad z \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{61}$$

it is obvious that:

$$H(z) = H\left(\log U - \log(1-U) + \log\left(\frac{\alpha_1}{\alpha_2}\right)\right) \tag{62}$$
$$\text{where } H \text{ is the unit step function}$$

$$
\begin{aligned}
\Pr\big([b \equiv H(z)] = 1\big) &= P(z_1 \geq z_2) \\
&= P(z_1 - z_2 \geq 0) \\
&= P\Bigg(\underbrace{\log U - \log(1-U) + \log\frac{\alpha_1}{\alpha_2}}_{z} \geq 0\Bigg) \\
&= P(z \geq 0)
\end{aligned}
\tag{63}
$$

in case $\alpha_1 = \theta$ and $\alpha_2 = 1-\theta$:

$$
z = \log\frac{\theta}{1-\theta} + \log\frac{U}{1-U} \tag{64}
$$

6.3.3 verify re-parameterization of u:

in here, we work backwards to verify:

$$
\begin{aligned}
&H(z) = 1 \quad \text{if } z > 0 \\
\implies\ & b = H(z(\theta,u)) = 1 \quad \text{if } z(\theta,u) > 0
\end{aligned}
\tag{65}
$$

so to find $\Pr(b=1)$ we just need to find $p\big(z(\theta,u) \geq 0\big)$, it turns out that $z(\theta,u)$ is monotonically increasing, so, let $u_0 \in (0,1)$

$$
\begin{aligned}
\Pr(b=1) \equiv \Pr\big(z(\theta,u) \geq 0\big) &= \Pr\Big(u \geq u_0 \big| \underbrace{z(\theta,u_0) = 0}\Big) \\
&= 1 - u_0 \qquad u_0 \text{ is some cut-off}
\end{aligned}
\tag{66}
$$

need to find cut-off u_0, such that $z(\theta,u_0) = 0$:

$$
\begin{aligned}
z(\theta,u_0) = 0 \implies & \log\left(\frac{\theta}{1-\theta}\right) + \log\left(\frac{u_0}{1-u_0}\right) = 0 \\
\implies & \log\left(\frac{u_0}{1-u_0}\right) = -\log\left(\frac{\theta}{1-\theta}\right) \implies \log\left(\frac{u_0}{1-u_0}\right) = \log\left(\frac{1-\theta}{\theta}\right) \\
\implies & \frac{u_0}{1-u_0} = \frac{1-\theta}{\theta} \\
\implies & u_0 = \frac{1-\theta}{\theta} \times (1-u_0) \implies u_0 \times \left(1 + \frac{1-\theta}{\theta}\right) = \frac{1-\theta}{\theta} \\
\implies & u_0 \times \frac{1}{\theta} = \frac{1-\theta}{\theta} \implies u_0 = 1-\theta
\end{aligned}
\tag{67}
$$

to summarize:

$$
\begin{aligned}
&\Pr(b=1|\theta) = \Pr(z(\theta,u) > 0|\theta) = \Pr(u > u_0) = 1 - u_0 = \theta \\
&\quad \text{where } u \sim U(0,1)
\end{aligned}
\tag{68}
$$

which shows it's a Bernoulli distribution

6.3.4 $p(z|b)$ under binary case

it may be tempting to solve it in the same way as you do for multi-class case, i.e., looking at the joint density and hope to deduce $p(z|b)$ of the something like:

$$\begin{aligned}
p(b_k = 1|z)p(z) &= \underbrace{\text{Logistic}(z; \log(\theta) - \log(1-\theta))}_{p(z)} \underbrace{\mathbf{1}(z \geq 0)}_{p(b=1|z)} \\
&= \frac{\exp\big(-(z-\mu)\big)}{\big(1+\exp\big(-(z-\mu)\big)\big)^2}\mathbf{1}(z \geq 0) \\
&= \frac{\exp\left(-\left(z - \log\theta + \log(1-\theta)\right)\right)}{\left(1+\exp\left(-\left(z - \log\theta + \log(1-\theta)\right)\right)\right)^2}\mathbf{1}(z \geq 0)
\end{aligned} \tag{69}$$

it is difficult to see where this leads, but there is an easier way to solve it, looking at [**?**]:

$$\begin{aligned}
b = 1 \implies z = \log\frac{\theta}{1-\theta} + \log\frac{u}{1-u} &\geq 0 \\
\implies \log\frac{u}{1-u} &\geq -\log\frac{\theta}{1-\theta} \\
&\geq \log\frac{1-\theta}{\theta} \\
\implies \frac{u}{1-u} &\geq \frac{1-\theta}{\theta} \\
\implies u\,\theta &\geq (1-u)(1-\theta) \qquad \text{valid as } \theta \geq 0 \quad u \geq 0 \\
&\geq 1 - u - \theta + u\theta \\
\implies u &\geq 1-\theta
\end{aligned} \tag{70}$$

in words, it meant the unit line is divided by the point $1-\theta$:

$$b = \begin{cases} 1 & u \in (1-\theta, 1) \qquad \text{this interval has length } \theta \\ 0 & u \in (0, 1-\theta) \qquad \text{this interval has length } 1-\theta \end{cases} \tag{71}$$

so instead of sampling $u \sim \mathcal{U}(0,1)$ in the prior case, for sampling $p(z|b)$ we transform u conditionally by:

$$b = \begin{cases} 1 & v' = (1-\theta) + \theta u \\ 0 & v' = (1-\theta)u \end{cases} \tag{72}$$

then compute the inverse of Logistic distribution using v':

$$z|b = \log\frac{\theta}{1-\theta} + \log\frac{v'}{1-v'} \tag{73}$$

7 Generalised Re-parameterization

7.1 a quick revision on change of variables in probablity

7.1.1 continous PDF

for re-parameterization knowing:

$$\begin{aligned} q_\epsilon(\epsilon) &= \left|\det \frac{\partial z}{\partial \epsilon}\right| q(z) \\ q_z(z) &= \left|\det \frac{\partial \epsilon}{\partial z}\right| q_\epsilon(\epsilon) \end{aligned} \tag{74}$$

because let $z = g(\epsilon) \implies \epsilon = g^{-1}(z)$:

$$\begin{aligned} \Pr_Z(z) &= \Pr_Z(Z \leq z) \\ &= \Pr_Z(g(\mathcal{E}) \leq z) \\ &= \Pr_{\mathcal{E}}(\mathcal{E} \leq g^{-1}(z)) \\ &= \Pr_{\mathcal{E}}(\mathcal{E} \leq \epsilon) \end{aligned} \tag{75}$$

$$\begin{aligned} p_Z(z) &= \frac{\mathrm{d}\,\Pr_Z(z)}{\mathrm{d}z} = \frac{\mathrm{d}\,\Pr_{\mathcal{E}}(\epsilon)}{\mathrm{d}z} \\ &= \frac{\mathrm{d}\Pr_{\mathcal{E}}(\epsilon)}{\mathrm{d}\epsilon}\frac{\mathrm{d}\epsilon}{\mathrm{d}z} \\ &= p_\epsilon(\epsilon)\frac{\mathrm{d}\epsilon}{\mathrm{d}z} \\ &= p_\epsilon(\epsilon)\left|\frac{\mathrm{d}\epsilon}{\mathrm{d}z}\right| \quad \text{because of probablity} \\ &\equiv p_\epsilon(\epsilon)\left|\det \frac{\partial \epsilon}{\partial z}\right| \quad \text{for multi-dimension} \end{aligned}$$

when computing relationship between $\mathrm{d}\epsilon$ and $\mathrm{d}z$:

$$\left|\det \frac{\partial \epsilon}{\partial z}\right| \mathrm{d}z = \mathrm{d}\epsilon \tag{76}$$

when computing expectation:

$$\begin{aligned} \mathbb{E}_{q(z)}[f(z)] &= \int_z f(z) q(z) \mathrm{d}z \\ &= \int_z f(z)\Big(\left|\det \frac{\partial \epsilon}{\partial z}\right| q_\epsilon(\epsilon)\Big)\mathrm{d}z \\ &= \int_\epsilon f(g(\epsilon)) q_\epsilon(\epsilon)\Big(\left|\det \frac{\partial \epsilon}{\partial z}\right|\mathrm{d}z\Big) \\ &= \int_\epsilon f(g(\epsilon)) q_\epsilon(\epsilon)\mathrm{d}\epsilon \end{aligned} \tag{77}$$

looking at the relationship:

$$\int_z f(z)q(z)\mathrm{d}z = \int_\epsilon f(g(\epsilon))q_\epsilon(\epsilon)\mathrm{d}\epsilon \tag{78}$$

one may find that the Jacobian does not appear in the above equation. Therefore, as long as one may **establish** relationship that:

$$q(\epsilon) = \big|\det\frac{\partial z}{\partial \epsilon}\big| q(z)$$

then, we can use use the Eq.(77) for that. Note that without $q(z) = \big|\det\frac{\partial \epsilon}{\partial z}\big|q(\epsilon)$ being established then equation Eq.(77) can not be true, as:

$$\begin{aligned}\mathbb{E}_{q(z)}[f(z)] &= \int_z f(z)q(z)\mathrm{d}z \\ &\neq \int_z f(z)\Big(\Big|\det\frac{\partial \epsilon}{\partial z}\Big|q_\epsilon(\epsilon)\Big)\mathrm{d}z \\ \implies \mathbb{E}_{q(z)}[f(z)] &\neq \int_\epsilon f(g(\epsilon))q_\epsilon(\epsilon)\mathrm{d}\epsilon\end{aligned} \tag{79}$$

For example, remember in Normalizing Flow, knowing the relationsip between $p(\mathbf{z}_k) = \Big|\det\frac{\partial \mathbf{z}_k}{\partial \mathbf{z}_0}\Big|p(\mathbf{z}_0)$ we can compute the expecation using samples from $\mathbf{z}_0$ only:

$$\begin{aligned}\mathbb{E}_{p(\mathbf{z}_K)}[h(\mathbf{z}_K)] &= \int_{\mathbf{z}_K} h(\mathbf{z}_K)p(\mathbf{z}_K)\mathrm{d}\mathbf{z}_K \\ &= \int_{\mathbf{z}_0} h\big(f_K\circ\cdots\circ f_2\circ f_1(\mathbf{z}_0)\big)p(\mathbf{z}_0)\mathrm{d}\mathbf{z}_0 \\ &= \mathbb{E}_{p(\mathbf{z}_0)}\big[h\big(f_K\circ\cdots\circ f_2\circ f_1(\mathbf{z}_0)\big)\big]\end{aligned} \tag{80}$$

Another example is the well known Gaussian case, we let:

$$p_\epsilon(\epsilon) = \frac{1}{\sqrt{2\pi}}\exp^{-\epsilon^2/2} \equiv \mathcal{N}(\epsilon;0,1) \tag{81}$$

and letting:

$$\begin{aligned}g(\theta,\epsilon) &= \mu+\sigma\epsilon \\ \implies \epsilon = g^{-1}(z,\theta) &= \frac{z-\mu}{\sigma} \\ \implies \Big|\frac{\mathrm{d}\epsilon}{\mathrm{d}z}\Big| &= \sigma^{-1}\end{aligned} \tag{82}$$

substituting formula:

$$
\begin{aligned}
p_Z(z) &= \left|\frac{\mathrm{d}\epsilon}{\mathrm{d}z}\right| p_\epsilon(\epsilon) \\
&= \sigma^{-1} p_\epsilon(g^{-1}(\theta, z)) \\
&= \sigma^{-1} \frac{1}{\sqrt{2\pi}} \exp^{-\left(\frac{z-\mu}{\sigma}\right)^2/2} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}
\end{aligned} \tag{83}
$$

7.1.2 discrete case

$$
p_Y(Y=y) = \sum_{x \in g^{-1}(y)} p_X(x) \tag{84}
$$

apply to Gumbel-max case:

$$
\begin{aligned}
p_K(K=k) &= \int_{\{z_i\}\,:\,\arg\max_i\left(\left\{z_i+\phi_i\right\}\right)=k} \mathrm{ga}_Z(\{z_i\}) \\
&= \frac{\exp(\phi_k)}{\sum_i \exp(\phi_i)}
\end{aligned} \tag{85}
$$

it's actually simpler, as there is no Jacobian stuff

7.2 independent case

coming back to our focus of re-parameterization in ∇_θ:

$$
\begin{aligned}
\nabla_\theta \mathbb{E}_{q(z;\theta)}[f(z)] &= \nabla_\theta \int_z f(z) q(z;\theta) \mathrm{d}z \\
&= \nabla_\theta \int_z f(z) \left|\det \frac{\partial \epsilon}{\partial z}\right| q_\epsilon(\epsilon) \mathrm{d}z \\
&= \int_z q_\epsilon(\epsilon) \nabla_\theta f(z) \left|\det \frac{\partial \epsilon}{\partial z}\right| \mathrm{d}z \\
&= \int_z \underbrace{q_\epsilon\big(g^{-1}(z)\big) \nabla_\theta f(z)}_{\phi(z)} \left|\det \frac{\partial \epsilon}{\partial z}\right| \mathrm{d}z
\end{aligned} \tag{86}
$$

using generic change of variable in integral, and assuming:

$$
\int_{z \in \mathcal{Z}} \phi(T(z)) \left|\det \frac{\partial \epsilon}{\partial z}\right| \mathrm{d}z = \int_{\epsilon \in T(\mathcal{Z})} \phi(\epsilon)\, \mathrm{d}\epsilon \tag{87}
$$

let $T(z) \to g^{-1}(z)$, the RHS of Eq.(86):

$$
\begin{aligned}
\nabla_\theta \mathbb{E}_{q(z;\theta)}[f(z)] &= \nabla_\theta \int_z f(z) q(z;\theta) \mathrm{d}z \\
&= \int_\epsilon \underbrace{q_\epsilon(\epsilon) \nabla_\theta f\big(g(\epsilon,\theta)\big)}_{\phi(\epsilon)} \mathrm{d}\epsilon \qquad \text{RHS of Eq.(87)}
\end{aligned} \tag{88}
$$

it does **not** work if one was attempt to change $q_\epsilon(\epsilon) \to q(z;\theta)$ back, i.e.,:

$$\int_z q(z;\theta)\nabla_\theta f(z)\mathrm{d}z \tag{89}$$

if you can do this, then ∇_θ has to be put in front of $q(z;\theta)$ to look like Eq.(86). Otherwise, conducting this "virtual" change of variable will make:

$$\int_z q(z;\theta)\nabla_\theta f(z)\mathrm{d}z \neq \int_z \nabla_\theta f(z)q(z;\theta)\mathrm{d}z \tag{90}$$

challenging conventional thinking, one usually think re-parameterization works by $\epsilon \sim q_\epsilon(\epsilon)$ first; however, one can replace this by:

$$z \sim q(z;\theta) \qquad \epsilon = g^{-1}(z,\theta) \tag{91}$$

one way to think about this is that the θ effect by $q(z;\theta)$ is "counteracted" by $g^{-1}(z,\theta)$, such as ϵ independent of θ

7.3 "weakly" dependent case

now look at a situation where ϵ and θ are "weakly" dependent, this means both ϵ and z are dependant on θ:

$$q_\epsilon(\epsilon;\theta) \longleftrightarrow q(z;\theta) \tag{92}$$

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{q(z;\theta)}[f(z)] &= \nabla_\theta \int_z f(z)q(z;\theta)\mathrm{d}z \\
&= \int_\epsilon \nabla_\theta \Big(\underbrace{f\big(g(\epsilon,\theta)\big)}_{u} \underbrace{q(\epsilon;\theta)}_{v} \Big)\mathrm{d}\epsilon \qquad \text{re-param } \epsilon \text{ using Eq.(88), but placing everything after } \nabla_\theta \\
&= \int_\epsilon q_\epsilon(\epsilon;\theta)\nabla_\theta f\big(g(\epsilon,\theta)\big)\mathrm{d}\epsilon + \int_\epsilon f\big(g(\epsilon,\theta)\big)\underline{\nabla_\theta q_\epsilon(\epsilon;\theta)}\mathrm{d}\epsilon \\
&= \underbrace{\int_\epsilon q_\epsilon(\epsilon;\theta)\nabla_\theta f\big(g(\epsilon,\theta)\big)\mathrm{d}\epsilon}_{g^{\text{rep}}} + \underbrace{\int_\epsilon f\big(g(\epsilon,\theta)\big)\underline{q_\epsilon(\epsilon;\theta)\nabla_\theta \log q_\epsilon(\epsilon;\theta)}\mathrm{d}\epsilon}_{g^{\text{corr}}} \qquad \text{Reinforce trick}
\end{aligned} \tag{93}$$

7.3.1 look at g^{rep}:

$$
\begin{aligned}
g^{\text{rep}} &= \int_\epsilon q_\epsilon(\epsilon,\theta)\nabla_\theta f\big(g(\epsilon,\theta)\big)\mathrm{d}\epsilon \\
&= \int_\epsilon \left|\det\frac{\partial z}{\partial \epsilon}\right| q(z;\theta)\underline{\nabla_\theta f\big(g(\epsilon,\theta)\big)}\mathrm{d}\epsilon \qquad \text{using Eq.(74)} \\
&= \int_e p\left|\det\frac{\partial z}{\partial \epsilon}\right| q(z;\theta)\underline{\nabla_z f(z)\big|_{z=g(\epsilon,\theta)}\nabla_\theta g(\epsilon,\theta)}\mathrm{d}\epsilon \qquad \text{using chain rule} \\
&= \int_\epsilon q(z;\theta)\nabla_z f(z) h(\epsilon,\theta)\left|\det\frac{\partial z}{\partial \epsilon}\right|\mathrm{d}\epsilon \\
&= \int_z q(z;\theta)\nabla_z f(z)\nabla_\theta g\big(g^{-1}(z,\theta),\theta\big)\mathrm{d}z \qquad \text{re-parameterization to } z \text{ via Eq.(87)} \\
&= \mathbb{E}_{q(z;\theta)}\Big[\nabla_z f(z)\nabla_\theta g\big(g^{-1}(z,\theta),\theta\big)\Big]
\end{aligned} \tag{94}
$$

7.3.2 look at g^{corr}:

$$
\text{let } u(\epsilon,\theta) \equiv \nabla_\theta \log\left|\det\frac{\partial z}{\partial \epsilon}\right| \tag{95}
$$

$$
\begin{aligned}
g^{\text{corr}} &= \int_\epsilon q_\epsilon(\epsilon;\theta) f\big(g(\epsilon,\theta)\big)\nabla_\theta \log q_\epsilon(\epsilon;\theta)\mathrm{d}\epsilon \\
&= \int_\epsilon \left|\det\frac{\partial z}{\partial \epsilon}\right| \underbrace{q(z;\theta) f(z)}\,\nabla_\theta \log\left(\left|\det\frac{\partial z}{\partial \epsilon}\right| q(z;\theta)\right)\mathrm{d}\epsilon \qquad \text{using Eq.(74)} \\
&= \int_\epsilon \underbrace{q(z;\theta) f(z)}\left(\nabla_\theta \log\left|\det\frac{\partial z}{\partial \epsilon}\right| + \nabla_\theta \log q\big(g(\epsilon,\theta);\theta\big)\right)\left|\det\frac{\partial z}{\partial \epsilon}\right|\mathrm{d}\epsilon
\end{aligned} \tag{96}
$$

looking at $\nabla_\theta \log q\big(g(\epsilon,\theta);\theta\big)$, there are two path of derivatives for computing ∇_θ:

$$
\nabla_\theta \log q\big(\underbrace{g(\epsilon,\theta)}_{1};\underbrace{\theta}_{2}\big) = \underbrace{\nabla_z \log q\big(z;\theta\big)\nabla_\theta g(\epsilon;\theta)}_{1} + \underbrace{\nabla_\theta \log q\big(z;\theta\big)}_{2} \tag{97}
$$

$$
\begin{aligned}
&= \int_z q(z;\theta) f(z)\left(\nabla_\theta \log\left|\det\frac{\partial z}{\partial \epsilon}\right| + \nabla_z \log q\big(z;\theta\big)\nabla_\theta g(\epsilon;\theta) + \nabla_\theta \log q\big(z;\theta\big)\right)\mathrm{d}z \quad \text{re-parameterize to } z \\
&= \int_z q(z;\theta) f(z)\Big(u(\epsilon,\theta) + \nabla_z \log q\big(z;\theta\big)\nabla_\theta g(\epsilon,\theta) + \nabla_\theta \log q\big(z;\theta\big)\Big)\mathrm{d}z \\
&= \int_z q(z;\theta) f(z)\Big(u(g^{-1}(z,\theta),\theta) + \nabla_z \log q\big(z;\theta\big)\nabla_\theta g\big(g^{-1}(z,\theta),\theta\big) + \nabla_\theta \log q\big(z;\theta\big)\Big)\mathrm{d}z \quad \text{replace all } \epsilon \to g^{-1}(\epsilon,\theta) \\
&= \mathbb{E}_{q(z;\theta)}\Big[f(z)\Big(u(g^{-1}(z,\theta),\theta) + \nabla_z \log q\big(z;\theta\big)\nabla_\theta g\big(g^{-1}(z,\theta),\theta\big) + \nabla_\theta \log q\big(z;\theta\big)\Big)\Big]
\end{aligned} \tag{98}
$$

7.3.3 to compute weakly-dependent re-parameterization

to summarize:

$$\begin{aligned}\nabla_\theta \mathbb{E}_{q(z;\theta)}[f(z)] &= g^{\text{rep}} + g^{\text{corr}} \\ &= \mathbb{E}_{q(z;\theta)}\Big[\nabla_z f(z) \nabla_\theta g\big(g^{-1}(z,\theta),\theta\big)\Big] \\ &\quad + \mathbb{E}_{q(z;\theta)}\Big[f(z)\Big(u(g^{-1}(z,\theta),\theta) + \nabla_z \log q\big(z;\theta\big)\nabla_\theta g\big(g^{-1}(z,\theta),\theta\big) + \nabla_\theta \log q\big(z;\theta\big)\Big)\Big]\end{aligned} \tag{99}$$

therefore, it seems that to re-parameterize any distribution $q(z;\theta)$, one needs to first compute:

1. $\nabla_\theta g\big(\epsilon,\theta\big) = \nabla_\theta g\big(g^{-1}(z,\theta),\theta\big)$
2. $u(\epsilon,\theta) \equiv \nabla_\theta \log \left|\det \frac{\partial z}{\partial \epsilon}\right|$

7.4 Example: Dirichlet Distribution

looking at definition of Dirichlet distribution:
$q(\mathbf{z};\boldsymbol{\alpha}) \equiv \text{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \equiv [\alpha_1,\ldots,\alpha_k]$, and $\boldsymbol{\alpha}_0 = \sum_i \alpha_i$

$$\begin{aligned}\mathbf{z} = g(\boldsymbol{\epsilon},\boldsymbol{\alpha}) &= \exp\left(\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\alpha})\,\boldsymbol{\epsilon} + \boldsymbol{\mu}(\boldsymbol{\alpha})\right) \\ = g(\boldsymbol{\epsilon},\boldsymbol{\alpha}) &= \exp\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\,\boldsymbol{\epsilon} + \boldsymbol{\mu}\right) \qquad \text{for short}\end{aligned} \tag{100}$$

and

$$\boldsymbol{\mu} = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\alpha})}[\log(\mathbf{z})] = \begin{bmatrix}\psi(\alpha_1) - \psi(\boldsymbol{\alpha}_0) \\ \vdots \\ \psi(\alpha_k) - \psi(\boldsymbol{\alpha}_0)\end{bmatrix} \tag{101}$$

looking at individual element of co-variance matrix $\Sigma_{i,j}$

$$\Sigma_{i,j} = \mathbf{Cov}(\log(z_i),\log(z_j)) = \begin{cases}\psi(\alpha_i) - \psi'(\boldsymbol{\alpha}_0) & \text{if } i=j \\ -\psi'(\boldsymbol{\alpha}_0) & \text{otherwise}\end{cases} \tag{102}$$

the whole $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix}\psi'(\alpha_1) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \psi'(\alpha_k)\end{bmatrix} - \psi'(\boldsymbol{\alpha}_0)\begin{bmatrix}1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1\end{bmatrix} \tag{103}$$

by construction, $\boldsymbol{\Sigma}$ is positive definite, hence:

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top \\ \implies \boldsymbol{\Sigma}^{\frac{1}{2}} &= \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{V}^\top\end{aligned} \tag{104}$$

let's look at:

$$
\begin{aligned}
J = \frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}} &= \frac{\partial \exp\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\epsilon} + \boldsymbol{\mu}\right)}{\partial \boldsymbol{\epsilon}} \\
&= \frac{\partial \left(\begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix} \right)}{\partial \boldsymbol{\epsilon}} = \frac{\partial \exp \left(\begin{bmatrix} \boldsymbol{\Sigma}^{\frac{1}{2}}{}_{1,:}\, \boldsymbol{\epsilon} + \mu_1 \\ \vdots \\ \boldsymbol{\Sigma}^{\frac{1}{2}}{}_{k,:} \boldsymbol{\epsilon} + \mu_k \end{bmatrix} \right)}{\partial \boldsymbol{\epsilon}} \\
\implies J_{i,j} &= \Sigma_{i,j}^{1/2} \exp\left(\Sigma^{\frac{1}{2}}{}_{i,:}{}^{\top} \boldsymbol{\epsilon} + \mu_i\right) \\
&= \Sigma_{i,j}^{1/2} z_i
\end{aligned}
\tag{105}
$$

it is then obvious that the Jacobian J looks like:

$$
\boldsymbol{J} = \begin{bmatrix} \Sigma_{1,1}^{1/2} z_1 & \dots \Sigma_{1,k}^{1/2} z_1 & \\ \vdots & \ddots & \vdots \\ \Sigma_{k,1}^{1/2} z_k & \dots \Sigma_{k,k}^{1/2} z_k & \end{bmatrix} = \underbrace{\begin{bmatrix} \Sigma_{1,1}^{1/2} z_1 & \dots \Sigma_{1,k}^{1/2} z_1 & \\ \vdots & \ddots & \vdots \\ \Sigma_{1,k}^{1/2} z_k & \dots \Sigma_{k,k}^{1/2} z_k & \end{bmatrix}}_{\text{co-variance matrix is symmetric}} = \boldsymbol{\Sigma}^{\frac{1}{2}} \begin{bmatrix} z_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & z_k \end{bmatrix}
\tag{106}
$$

substitute J in:

$$
\begin{aligned}
\left| \det \frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}} \right| = \left| \det J \right| &= \left| \det \left(\boldsymbol{\Sigma}^{\frac{1}{2}} \begin{bmatrix} z_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & z_k \end{bmatrix} \right) \right| \\
&= \det\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right) \prod_{i=1}^{k} z_i \qquad \text{drop absolute operator as all positive}
\end{aligned}
\tag{107}
$$

from definition of Dirichlet distribution:

$$
\begin{aligned}
\mathrm{Dir}(\mathbf{z}; \boldsymbol{\alpha}) &= \frac{\Gamma(\boldsymbol{\alpha}_0)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} z_i^{\alpha_i - 1} \\
\implies \log\left(\mathrm{Dir}(\mathbf{z}; \boldsymbol{\alpha})\right) &= \log\left(\Gamma(\boldsymbol{\alpha}_0)\right) - \sum_{i}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1) \log(z_i)
\end{aligned}
\tag{108}
$$

derivative of $\log q(\mathbf{z}; \boldsymbol{\alpha})$ is:

$$
\begin{aligned}
\frac{\partial}{\partial z_i} \log q(\mathbf{z}; \boldsymbol{\alpha}) &= \frac{\alpha_i - 1}{z_i} \\
\frac{\partial}{\partial \alpha_i} \log q(\mathbf{z}; \boldsymbol{\alpha}) &= \frac{\partial \log \Gamma(\boldsymbol{\alpha}_0)}{\partial \alpha_i} + \frac{\partial \log \Gamma(\alpha_i)}{\partial \alpha_i} + \frac{\partial \alpha_i \log(z_i)}{\partial \alpha_i} \\
&= \psi(\boldsymbol{\alpha}_0) \frac{\partial \boldsymbol{\alpha}_0}{\partial \alpha_i} + \psi(\alpha_i) + \log(z_i) \\
&= \psi(\boldsymbol{\alpha}_0) + \psi(\alpha_i) + \log(z_i)
\end{aligned}
\tag{109}
$$

7.4.1 Finding $\nabla_\theta g(\epsilon, \theta)$ and $u(\epsilon, \theta) \equiv \nabla_\theta \log \left| \det \frac{\partial z}{\partial \epsilon} \right|$ for Dirichlet

and be reminded again that:

$$\mathbf{z} = g(\boldsymbol{\epsilon}, \boldsymbol{\alpha}) = \exp\left(\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\alpha})\,\boldsymbol{\epsilon} + \boldsymbol{\mu}(\boldsymbol{\alpha})\right) \tag{110}$$

for $\nabla_\theta g(\epsilon, \theta)$, looking at individual Jacobian matrix: $\frac{\partial z_i}{\alpha_j}$:

$$\begin{aligned}
\frac{\partial z_i}{\alpha_j} &= \frac{\exp\left(\Sigma^{\frac{1}{2}}{}_{i,:}\boldsymbol{\epsilon} + \mu_i\right)}{\partial\left(\Sigma^{\frac{1}{2}}{}_{i,:}\boldsymbol{\epsilon} + \mu_i\right)} \, \frac{\partial\left(\Sigma^{\frac{1}{2}}{}_{i,:}\boldsymbol{\epsilon} + \mu_i\right)}{\partial \alpha_j} \\
&= z_i \, \frac{\partial\left(\Sigma^{\frac{1}{2}}{}_{i,:}\boldsymbol{\epsilon} + \mu_i\right)}{\partial \alpha_j} \\
&= z_i(\boldsymbol{\epsilon}, \boldsymbol{\alpha}) \, \frac{\partial\left(\Sigma^{\frac{1}{2}}{}_{i,:}(\boldsymbol{\alpha})\boldsymbol{\epsilon} + \mu_i(\boldsymbol{\alpha})\right)}{\partial \alpha_j} \qquad \text{add } \boldsymbol{\alpha} \\
&= z_i(\boldsymbol{\epsilon}, \boldsymbol{\alpha}) \, \frac{\partial \Sigma^{\frac{1}{2}}{}_{i,:}(\boldsymbol{\alpha})}{\partial \alpha_j}\boldsymbol{\epsilon} + \frac{\partial \mu_i(\boldsymbol{\alpha})}{\partial \alpha_j}
\end{aligned} \tag{111}$$

fill the whole Jocobian: $\frac{\partial \mathbf{z}}{\boldsymbol{\alpha}}$:

$$\begin{aligned}
\nabla_\theta g(\epsilon, \theta) &= \begin{bmatrix} \frac{\partial z_1}{\partial \alpha_1} & \cdots & \frac{\partial z_1}{\partial \alpha_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_k}{\partial \alpha_1} & \cdots & \frac{\partial z_k}{\partial \alpha_k} \end{bmatrix} \\
&= \begin{bmatrix} z_1(\boldsymbol{\epsilon}, \boldsymbol{\alpha}) \, \frac{\partial \Sigma^{\frac{1}{2}}{}_{1,:}(\boldsymbol{\alpha})}{\partial \alpha_1}\boldsymbol{\epsilon} + \frac{\partial \mu_1(\boldsymbol{\alpha})}{\partial \alpha_1} & \cdots & z_1(\boldsymbol{\epsilon}, \boldsymbol{\alpha}) \, \frac{\partial \Sigma^{\frac{1}{2}}{}_{1,:}(\boldsymbol{\alpha})}{\partial \alpha_k}\boldsymbol{\epsilon} + \frac{\partial \mu_1(\boldsymbol{\alpha})}{\partial \alpha_k} \\ \vdots & \ddots & \vdots \\ z_k(\boldsymbol{\epsilon}, \boldsymbol{\alpha}) \, \frac{\partial \Sigma^{\frac{1}{2}}{}_{k,:}(\boldsymbol{\alpha})}{\partial \alpha_1}\boldsymbol{\epsilon} + \frac{\partial \mu_k(\boldsymbol{\alpha})}{\partial \alpha_1} & \cdots & z_k(\boldsymbol{\epsilon}, \boldsymbol{\alpha}) \, \frac{\partial \Sigma^{\frac{1}{2}}{}_{k,:}(\boldsymbol{\alpha})}{\partial \alpha_k}\boldsymbol{\epsilon} + \frac{\partial \mu_k(\boldsymbol{\alpha})}{\partial \alpha_k} \end{bmatrix}
\end{aligned} \tag{112}$$

to find $u(\epsilon, \theta) \equiv \nabla_\theta \log \left| \det \frac{\partial z}{\partial \epsilon} \right|$, since we know

$$\begin{aligned}
\left| \det \frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}} \right| &= \det\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right) \prod_{i=1}^{k} z_i \\
\implies \nabla_{\boldsymbol{\alpha}} \log \left| \det \frac{\partial \mathbf{z}}{\partial \epsilon} \right| &= \nabla_{\boldsymbol{\alpha}} \log \left(\det\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right) \prod_{i=1}^{k} \exp\left(\Sigma^{\frac{1}{2}}{}_{i,:}\boldsymbol{\epsilon} + \mu_i\right) \right) \\
&= \nabla_{\boldsymbol{\alpha}} \left[\log \det(\boldsymbol{\Sigma}^{\frac{1}{2}}) + \sum_{i=1}^{k} \Sigma^{\frac{1}{2}}{}_{i,:}\boldsymbol{\epsilon} + \mu_i \right]
\end{aligned} \tag{113}$$

note that $\nabla_{\boldsymbol{\alpha}} \log \left| \det \frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}} \right|$ is a vector, not a matrix:

$$\nabla_{\boldsymbol{\alpha}} \log \left| \det \frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}} \right| = \begin{bmatrix} \frac{\partial \log \det(\boldsymbol{\Sigma}^{\frac{1}{2}})}{\partial \alpha_1} + \frac{\partial \sum_{i=1}^{k} \Sigma^{\frac{1}{2}}(\boldsymbol{\alpha})_{i,:}\,\boldsymbol{\epsilon} + \mu_i(\boldsymbol{\alpha})}{\partial \alpha_1} \\ \vdots \\ \frac{\partial \log \det(\boldsymbol{\Sigma}^{\frac{1}{2}})}{\partial \alpha_k} + \frac{\partial \sum_{i=1}^{k} \Sigma^{\frac{1}{2}}(\boldsymbol{\alpha})_{i,:}\,\boldsymbol{\epsilon} + \mu_i(\boldsymbol{\alpha})}{\partial \alpha_k} \end{bmatrix} \tag{114}$$

so the only thing remain is to compute in $\nabla_\theta g\big(\epsilon,\theta\big)$ and $\nabla_{\boldsymbol{\alpha}} \log \Big| \det \frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}} \Big|$ is:

1. $\frac{\partial \boldsymbol{\mu}}{\partial \alpha_i}$: since $\boldsymbol{\mu} = \begin{bmatrix} \psi(\alpha_1) - \psi(\boldsymbol{\alpha}_0) \\ \vdots \\ \psi(\alpha_k) - \psi(\boldsymbol{\alpha}_0) \end{bmatrix}$, therefore:

$$\frac{\partial \boldsymbol{\mu}}{\partial \alpha_i} = \begin{bmatrix} -\frac{\partial \psi(\boldsymbol{\alpha}_0)}{\partial \alpha_i} \\ \vdots \\ \frac{\psi(\alpha_i)}{\partial \alpha_i} - \frac{\partial \psi(\boldsymbol{\alpha}_0)}{\partial \alpha_i} \\ \vdots \\ -\frac{\partial \psi(\boldsymbol{\alpha}_0)}{\partial \alpha_i} \end{bmatrix} = \begin{bmatrix} -\frac{\partial \psi(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0} \frac{\partial \boldsymbol{\alpha}_0}{\partial \alpha_i} \\ \vdots \\ \frac{\psi(\alpha_i)}{\partial \alpha_i} - \frac{\partial \psi(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0} \frac{\partial \boldsymbol{\alpha}_0}{\partial \alpha_i} \\ \vdots \\ -\frac{\partial \psi(\boldsymbol{\alpha}_0)}{\partial \boldsymbol{\alpha}_0} \frac{\partial \boldsymbol{\alpha}_0}{\partial \alpha_i} \end{bmatrix} = \begin{bmatrix} -\psi'(\boldsymbol{\alpha}_0) \\ \vdots \\ \psi'(\alpha_i) - \psi'(\boldsymbol{\alpha}_0) \\ \vdots \\ -\psi'(\boldsymbol{\alpha}_0) \end{bmatrix} \tag{115}$$

2. $\frac{\partial \boldsymbol{\Sigma}^{\frac{1}{2}}}{\alpha_i}$, this is hard, and from the original paper stating, $\frac{\partial \boldsymbol{\Sigma}^{\frac{1}{2}}}{\alpha_i}$ is the solution of Lyapunov equation:

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \alpha_i} = \frac{\partial \boldsymbol{\Sigma}^{\frac{1}{2}}}{\partial \alpha_i} \boldsymbol{\Sigma}^{\frac{1}{2}} + \boldsymbol{\Sigma}^{\frac{1}{2}} \frac{\partial \boldsymbol{\Sigma}^{\frac{1}{2}}}{\partial \alpha_i} \tag{116}$$

where, by change $\boldsymbol{\Sigma} \to \frac{\partial \boldsymbol{\Sigma}}{\partial \alpha_i}$ in Eq.(103):

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \alpha_i} = \begin{bmatrix} \psi''(\alpha_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \psi''(\alpha_k) \end{bmatrix} - \psi''(\boldsymbol{\alpha}_0) \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix} \tag{117}$$

3. $\frac{\partial \log \det(\boldsymbol{\Sigma}^{\frac{1}{2}})}{\partial \alpha_i}$:

this is easy and follow matrix cookbook, one may obtain:

$$\frac{\partial \log \det(\boldsymbol{\Sigma}^{\frac{1}{2}})}{\partial \alpha_i} = \mathrm{trace}\Big(\boldsymbol{\Sigma}^{\frac{1}{2}} \frac{\partial \boldsymbol{\Sigma}^{\frac{1}{2}}}{\partial \alpha_i}\Big) \tag{118}$$

7.5 Apply to Variational equation

$$\begin{aligned} \mathcal{L}(\theta) &= \mathbb{E}_{q(z|\theta)}\big[\log p(x,z) - \log q(z;\theta)\big] \\ &= \mathbb{E}_{q(z|\theta)}\big[\underbrace{\log p(x,z)}_{f(z)}\big] \underbrace{-\mathbb{E}_{q(z;\theta)}\big[\log q(z;\theta)\big]}_{H\big[q(z;\theta)\big]} \\ &= \mathbb{E}_{q(z|\theta)}\big[f(z)\big] + H\big[q(z;\theta)\big] \\ \nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \mathbb{E}_{q(z;\theta)}\big[f(z)\big] + \nabla_\theta H\big[q(z;\theta)\big] \end{aligned} \tag{119}$$

$$
\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \mathbb{E}_{q(z|\theta)}\big[f(z)\big] + \nabla_\theta H\big[q(z|\theta)\big] \\
&= g^{\text{rep}} + g^{\text{corr}} + \nabla_\theta H\big[q(z|\theta)\big] \\
&= \mathbb{E}_{q(z|\theta)}\Big[\nabla_z f(z) h\big(g^{-1}(z,\theta),\theta\big)\Big] \\
&\quad + \mathbb{E}_{q(z|\theta)}\Big[f(z)\Big(u(g^{-1}(z,\theta),\theta) + \nabla_z \log q\big(z|\theta\big) h\big(g^{-1}(z,\theta),\theta\big)\Big)\Big] \\
&\quad + \nabla_\theta H\big[q(z|\theta)\big]
\end{aligned} \tag{120}
$$

special case 1 : if distribution $q(\epsilon, \theta)$ does not depend on the variational parameters θ, then:

$$
\begin{aligned}
&\nabla_\theta \log q_\epsilon(\epsilon, \theta) = 0 \\
&\implies \nabla_\theta \mathcal{L}(\theta) = \int q_\epsilon(\epsilon,\theta) \nabla_\theta f\big(g(\epsilon,\theta)\big) \mathrm{d}\epsilon + \nabla_\theta H\big[q(z|\theta)\big]
\end{aligned} \tag{121}
$$

special case 2 : if $g(\epsilon, \theta) = \epsilon$:

$$
\begin{aligned}
g(\epsilon,\theta) = \epsilon &\implies \nabla_\theta g(\epsilon,\theta) = \nabla_\theta \epsilon = 0 \\
&\implies u(\epsilon,\theta) \equiv \nabla_\theta \log \big|\det \nabla_\epsilon g(\epsilon,\theta)\big| = \nabla_\theta \log \big|\det \nabla_\epsilon \epsilon\big| = 0
\end{aligned} \tag{122}
$$

$$
\begin{aligned}
\implies \nabla_\theta \mathcal{L}(\theta) &= g^{\text{rep}} + g^{\text{corr}} + H[q(z|\theta)] \\
&= \mathbb{E}_{q(z|\theta)}\Big[\nabla_z f(z) \underbrace{\nabla_\theta g\big(g^{-1}(z,\theta),\theta\big)}_{=0}\Big] \\
&\quad + \mathbb{E}_{q(z;\theta)}\Big[f(z)\Big(\underbrace{u(g^{-1}(z,\theta),\theta)}_{=0} + \nabla_z \log q\big(z;\theta\big) \underbrace{\nabla_\theta g\big(g^{-1}(z,\theta),\theta\big)}_{=0} + \nabla_\theta \log q\big(z;\theta\big)\Big)\Big] \\
&\quad + \nabla_\theta H\big[q(z|\theta)\big] \\
&= \mathbb{E}_{q(z|\theta)}\big[f(z) \nabla_\theta \log q\big(z;\theta\big)\big] + \nabla_\theta H\big[q(z|\theta)\big]
\end{aligned} \tag{123}
$$