1 Short Notes on Variational Bounds with Rescaled Terms

Learning a generative model of the form $p(x)=\int\ dzp(x,z)$ for $x\in\mathbb{R}^d$ and $z\in\mathbb{R}^m$ involves maximizing the likelihood of the data $\mathcal{L}=\sum_{x\in\text{Data}}\ln p(x)$.

When using variational approximations, we instead optimize the variational bound $\mathcal{F}=\sum_{x\in\text{Data}}\mathcal{F}(x)$, where

$$\mathcal{F}(x)\ =-\mathbb{E}_{q(z|x)}[\ln p(x|z)]+\text{KL}(q(z|x)\|p(z))\geqslant-\ln p(x), \tag{1}$$

and $q(z|x)$ is an arbitrary density with the same support as the marginal $p(z)$.

A nice property of the bound (1) is that it saturates when $q(z|x)$ is the true posterior distribution $p(z|x)$.

This can be seen by setting the functional derivative of (1) with respect to $q(z|x)$ to zero, taking into account the constraint $\int\ dzq(z|x)=1\forall x$ (with Lagrange multiplier $\lambda(x)$):

$$\begin{aligned}
\frac{\delta}{\delta q(z|x)}\left[\mathcal{F}(x)-\lambda(x)\int dzq(z|x)\right]\ &=-\ln p(x|z)+\ln q(z|x)-\ln p(z)+1-\lambda(x)\ =0\\
q^\star(z|x)\ &=p(x,z)e^{\lambda(x)}\\
\lambda(x)\ &=-\ln p(x)\\
&\Downarrow\\
q^\star(z|x)\ &=p(z|x).
\end{aligned}$$

Interestingly, the optimal Lagrange multiplier $\lambda(x)$ is the negative likelihood $\lambda(x)=-\ln p(x)$.

Additionaly note that if we replace the exact posterior in $\mathcal{F}(x)$ we also get $\mathcal{F}(x)=\lambda(x)$.

A couple of recent works, [3, 2, 4], explore modified variational losses $\mathcal{F}_\beta(x)$ [2, 4] and $\mathcal{F}_\kappa(x)$[1, 3] with rescaled KL or entropy terms of the form

$$\begin{aligned}
\mathcal{F}_\beta(x)\ &=-\mathbb{E}_{q(z|x)}[\ln p(x|z)]+\beta\text{KL}(q(z|x)\|p(z)), &(2)\\
\mathcal{F}_\kappa(x)\ &=-\mathbb{E}_{q(z|x)}[\ln p(x,z)]+\kappa\mathbb{E}_{q(z|x)}[\ln q(z|x)], &(3)
\end{aligned}$$

where $\beta,\kappa\geqslant 0$ are rescaling factors.

The loss (3) is also a key idea behind deterministic annealing methods [1, 3].

Inspired by these examples, I briefly discuss below a few properties of rescaled variational objectives.

1.1 Modified Bayes rule

The first point to note is that the optimal varitional densities $q_\beta^\star(z|x)$ and $q_\kappa^\star(z|x)$ for the losses $\mathcal{F}_\beta(x)$ and $\mathcal{F}_\kappa(x)$ respectively are no longuer the true posterior $p(z|x)$. Following the same derivation as before, we conclude that

$$\begin{aligned}
q_\beta^\star(z|x)\ &\propto p(z)e^{\frac{1}{\beta}\ln p(x|z)}\\
&\text{and}\\
q_\kappa^\star(z|x)\ &\propto e^{\frac{1}{\kappa}\ln p(x,z)}.
\end{aligned}$$

The optimal variational densities $q^{\star}_{\beta}(z|x)$ and $q^{\star}_{\kappa}(z|x)$ have slightly different interpretations, but can be very different on practice. The former changes the importance of the likelihood term while keeping the importance of the prior term fixed. The later changes the importance of both the likelihood and prior terms jointly. A consequence of this is that $q^{\star}_{\beta}(z|x)$ will interpolate between the prior $(\beta = \infty)$ and the maximum likelihood $(\beta = 0)$ as illustrated in Figure 1(middle). While $q^{\star}_{\kappa}(z|x)$ will interpolate between a flat density $(\kappa = \infty)$ and the MAP $(\kappa = 0)$ as illustrated in Figure 1(right).

In both cases the densities $q^{\star}_{\beta}(z|x)$ and $q^{\star}_{\kappa}(z|x)$ can be seen as a form of distorted Bayes rule for the generative model $p(x, z)$.

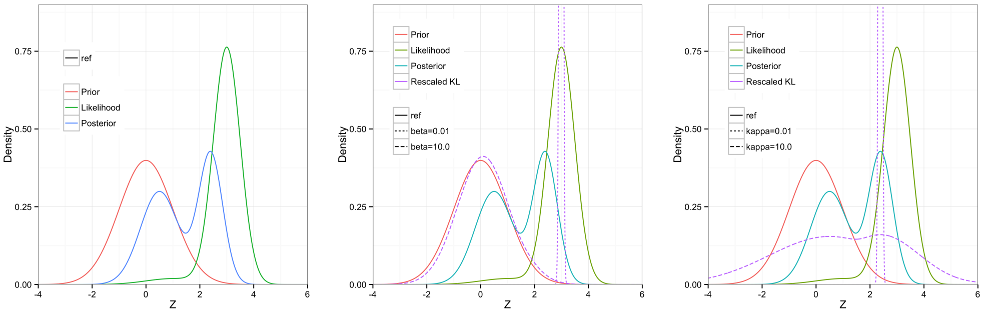

Figure 1. Illustration of differences between $q^{\star}(z|x)$, $q^{\star}_{\beta}(z|x)$ and $q^{\star}_{\kappa}(z|x)$. Left: Exact prior, (normalized) Likelihood and Posterior $q^{\star}(z|x)$; Middle: $q^{\star}_{\beta=0.01}(z|x)$ (dotted line) and $q^{\star}_{\beta=10}(z|x)$(dashed line); Right: $q^{\star}_{\kappa=0.01}(z|x)$ (dotted line) and $q^{\star}_{\kappa=10}(z|x)$(dashed line).

1.2 Bounds

The losses $\mathcal{F}_{\beta}(x)$ and $\mathcal{F}_{\kappa}(x)$ are typically not meant to be used with fixed β and κ, except perhaps in [4]. For instance, in deterministic annealing [1, 3] κ will typically converge to 1 during optimization with some schedule.

Nevertheless, if we want to use $\mathcal{F}_{\beta}(x)$ and $\mathcal{F}_{\kappa}(x)$ as alternative proxies for $-\ln\, p(x)$, they must form strict bounds. But there is no guarantee that the modified losses $\mathcal{F}_{\beta}(x)$ and $\mathcal{F}_{\kappa}(x)$ will correspond to bounds on $-\ln p(x)$ for $\beta, \kappa \neq 1$.

A trivial observation for $\mathcal{F}_{\beta}(x)$ is that since $\mathrm{KL}(q(z|x)\|p(z)) \geqslant 0$ we have

$$\mathcal{F}_{\beta}(x) \geqslant \mathcal{F}(x) \geqslant -\ln p(x) \quad \forall \beta \geqslant 1.$$

That is, for a fixed $q(z|x)$ and for any $\beta \geqslant 1$, $\mathcal{F}_{\beta}(x)$ is a valid bound but it is strictly worse than $\mathcal{F}(x)$. A more general bound for fixed $q(z|x)$ can be found with some algebra:

$$\begin{aligned}
\mathcal{F}(x) &= -\ln p(x) + \mathrm{KL}(q(z|x)\|p(z|x)) \\
\mathcal{F}_{\beta}(x) &= \mathcal{F}(x) + (\beta - 1)\mathrm{KL}(q(z|x)\|p(z)) \\
&= -\ln p(x) + \mathrm{KL}(q(z|x)\|p(z|x)) + (\beta - 1)\mathrm{KL}(q(z|x)\|p(z)) \\
&\Downarrow \\
&\mathrm{KL}(q(z|x)\|p(z|x)) + (\beta - 1)\mathrm{KL}(q(z|x)\|p(z)) \geqslant 0 \\
&\Downarrow \\
&\beta \geqslant 1 - \frac{\mathrm{KL}(q(z|x)\|p(z|x))}{\mathrm{KL}(q(z|x)\|p(z))}.
\end{aligned}$$

On practice it will be hard to use this bound as it requires knowledge of $\mathrm{KL}(q(z|x)\|p(z|x))$.

Similarly, we can obtain a bound on κ for $\mathcal{F}_\kappa(x)$:

$$\begin{aligned}\mathcal{F}_\kappa(x)+\ln p(x) &= \mathrm{KL}(q(z|x)\|p(z|x))+(\kappa-1)\mathbb{E}_{q(z|x)}[\ln q(z|x)]\geqslant 0\\ &\Downarrow\\ &\kappa\mathbb{E}_{q(z|x)}[\ln q(z|x)]\geqslant\mathbb{E}_{q(z|x)}[\ln q(z|x)]-\mathrm{KL}(q(z|x)\|p(z|x))\\ &\Updownarrow\\ &\begin{cases}\kappa\geqslant 1-\dfrac{\mathrm{KL}(q(z|x)\|p(z|x))}{\mathbb{E}_{q(z|x)}[\ln q(z|x)]} & \text{if } \mathbb{E}_{q(z|x)}[\ln q(z|x)]>0\\ \text{and} & \\ \kappa\leqslant 1-\dfrac{\mathrm{KL}(q(z|x)\|p(z|x))}{\mathbb{E}_{q(z|x)}[\ln q(z|x)]} & \text{if } \mathbb{E}_{q(z|x)}[\ln q(z|x)]<0\end{cases}.\end{aligned}$$

That is, whether κ must be bounded from above or from below will depend on the sign of the entropy term.

1.3 Relation to Up(Down)-Sampling

Below we assume that the elements of x are iid given z,

$$p(x|z) = \prod_{i=1}^{d} p_i(x_i|z).$$

Under this assumption, the bound (1) becomes

$$\mathcal{F}(x) = -\sum_i \mathbb{E}_{q(z|x)}[\ln p_i(x_i|z)]+\mathrm{KL}(q(z|x)\|p(z)). \tag{4}$$

Now imagine that we preprocess the data by replicating every element of x n times, $x_i\to(x_i,\ldots,x_i)$, while also replicating the densities $p_i(x_i|z)$. So that, $p_i((x_i,\ldots,x_i)|z)=p_i(x_i|z)^n$.

The corresponding variational bound $\mathcal{F}_n(x)$ for this modified problem is now given by

$$\mathcal{F}_n(x) = -n\sum_i \mathbb{E}_{q(z|x)}[\ln p_i(x_i|z)]+\mathrm{KL}(q(z|x)\|p(z)). \tag{5}$$

The optimal variational distribution $q_n^\star(z|x)$ can be computed as before and it gives

$$q_n^\star(z|x) \propto p(z)e^{n\ln p(x|z)}. \tag{6}$$

Interestingly, $q_n^\star(z|x)$ is equivalent to $q_\beta^\star(z|x)$ with $\beta=\frac{1}{n}$. That is, computing the variational posterior for an up-sampled image as above is mathematically equivalent to computing the variational posterior with a rescaled KL term when $\beta<1$.

The relation to down-sampling is much less evident as we must "erase" information from the model in a consitent manner. This can be done through a combination of a pooling operation and marginalization. The idea is to pool the data so that it becomes n times smaller, define a consistent density at the down-sampled image and up-sample it back. There is no single way of doing these operations.

Suppose we have a "pooling operator" f that takes the lth patch of n elements $(x_{i_1^l},\ldots,x_{i_n^l})$ and maps it to a pooled value $\hat{x}_l$, $f(x_{i_1^l},\ldots,x_{i_n^l})=\hat{x}_l$ (e.g. by averaging). Now, the joint density $p(x_{i_1},\ldots,x_{i_n}|z)$ will induce a density $\hat{p}(\hat{x}_l|z)$ by marginalizing over all values of $(x_{i_1^l},\ldots,x_{i_n^l})$ consistent with the same value of $\hat{x}_l$,

$$\hat{p_l}(\hat{x}_l|z) = \int\prod_{j=1}^{n} dx_{i_j^l}\, p(x_{i_1^l},\ldots,x_{i_n^l}|z)\delta(\hat{x}_l-f(x_{i_1^l},\ldots,x_{i_n^l})),$$

where $\delta(z)$ is the Dirac delta.

After pooling and marginalization, we now have a model defined on pooled (down-sampled) data. The last step is to up-sample back the data to its original size using the same up-sampling strategy as before (by replicating $\hat{x}_l \rightarrow (\hat{x}_{i_1^l} = \hat{x}_l, \ldots, \hat{x}_{i_n^l} = \hat{x}_l))$. This results in the relation $\hat{p}(\hat{x}_l|z) = \hat{p}((\hat{x}_{i_1^l}, \ldots, \hat{x}_{i_n^l})|z)^{\frac{1}{n}}$ and consequently $\hat{q}_n^{\star}(z|\hat{x}) \propto p(z) e^{\frac{1}{n} \ln \hat{p}\left(\left(\hat{x}_{i_1^l}, \ldots, \hat{x}_{i_n^l}\right)|z\right)}$. When the original (non-pooled) data and model are both smooth (e.g. very small variability inside a patch) we have that $\hat{p}((\hat{x}_{i_1^l}, \ldots, \hat{x}_{i_n^l})|z) \approx p(x_{i_1^l}, \ldots, x_{i_n^l}|z)$. So the relation between $q_{\beta}^{\star}(z|x)$ for $\beta > 1$ and down-sampling is only approximate.

These results suggest that we could see a tradeoff between spatial resolution and different values of β. Indeed such relationship has been observed in [2], as illustrated in figure (2).

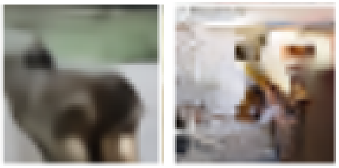

Figure 2. Tradeoff between spatial resolution and different values of β. Images from [2]. Left: $\beta > 1$ (model focuses on lower spatial frequencies); Right: $\beta < 1$ (model introduces higher-frequency details).

Bibliography

[1] Farhan Abrol, Stephan Mandt, Rajesh Ranganath, and David Blei. Deterministic annealing for stochastic variational inference. *stat*, 1050:7, 2014.

[2] Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra. Towards conceptual compression. *arXiv preprint arXiv:1604.08772*, 2016.

[3] San Gultekin, Aonan Zhang, and John Paisley. Stochastic annealing for variational inference. *arXiv preprint arXiv:1505.06723*, 2015.

[4] Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*, 2016.