0.0.1 General Setting of Variational Learning Framework

Principle 0.1 (Variational Learning Framework)**.** *Let $\mathcal{X}$ be the sample space $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ data points in $\mathcal{X}$. The general framework of* variational inference *consists of:*

1. *a* statistical model*: p_θ on $\mathcal{X}$ parameterized by $\theta \in \Theta$,*
2. *a* loss function*: $L(\mathcal{D}|p_\theta)$ that measures the goodness of the fit between model p_θ and the data $\mathcal{D}$, and often is of the form*
$$L(\mathcal{D}|p_\theta) = \sum_{x \in \mathcal{D}} \ell(x|p_\theta),$$
3. *a set $\mathcal{Q}$ of* admissible distributions *q over Θ, which is usually given by computational restrictions or simplifying assumptions,*
4. *a* regularization functional *D with arguments $q \in \mathcal{Q}$ that is supposed to penalize the complexity of p_θ, reflect model uncertainty, and incorporate prior knowledge and outcome expectations.*

Given all these choices one then solves the following optimization problem:

$$q_{\mathcal{D}} = \operatorname*{argmin}_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{q(\theta)} \left[L(\mathcal{D}|p_\theta) \right] + D(q) \right\}. \tag{1}$$

Prediction is then done via model averaging:

$$p(x|\mathcal{D}) := \int p_\theta(x)\, q_{\mathcal{D}}(d\theta). \tag{2}$$

Example 0.2. *We have the following examples as corner cases:*

1. Standard Variational Inference*:*

   $L(\mathcal{D}|p_\theta) := -\log p_\theta(\mathcal{D})$, $D(q) := \mathrm{KL}(q||\pi)$ with some prior π on Θ and $\mathcal{Q}$ some admissible distribution class.

2. Maximum likelihood estimation (MLE) *as variational inference:*

   $L(\mathcal{D}|p_\theta) := -\log p_\theta(\mathcal{D})$, $D := 0$, $\mathcal{Q} := \{\delta_{\tilde{\theta}} \,|\, \tilde{\theta} \in \Theta\}$ point-masses. Then we recover the maximum likelihood point estimator (written as a measure):
$$q_{\mathcal{D}}(\theta) = \delta_{\hat{\theta}_{MLE}}(\theta),$$
   with predictive distribution:
$$p_{\hat{\theta}_{MLE}}(x).$$

3. Maximum a-posteriori estimation (MAP) *as variational inference:*

Let π *be a* prior *distribution over* Θ *and put:* $L(\mathcal{D}|p_\theta) := -\log p_\theta(\mathcal{D})$, $D(q) := \mathrm{CE}(q||\pi)$, $\mathcal{Q} := \{\delta_{\tilde{\theta}} \,|\, \tilde{\theta} \in \Theta\}$ *point-masses. Then we recover the maximum a-posteriori point estimator (written as a measure):*

$$q_{\mathcal{D}}(\theta) = \delta_{\hat{\theta}_{MAP}}(\theta),$$

with predictive distribution:

$$p_{\hat{\theta}_{MAP}}(x).$$

4. Full Bayesian approach *as variational inference:*

Let π *be a* prior *distribution over* Θ *and put:* $L(\mathcal{D}|p_\theta) := -\log p_\theta(\mathcal{D})$, $D(q) := \mathrm{KL}(q||\pi)$, $\mathcal{Q} := \mathcal{P}(\Theta)$ *all probability measures. Then we recover the usual Bayesian posterior:*

$$q_{\mathcal{D}}(\theta) = \pi(\theta|\mathcal{D}),$$

with the usual predictive distribution:

$$p(x|\mathcal{D}) = \int p_\theta(x)\, \pi(d\theta|\mathcal{D}).$$

5. Generalized Bayes: *Use* $D(q) := \frac{1}{\beta}\,\mathrm{KL}(q||\pi)$ *with* $\beta > 0$.

6. Variational Bayes*:* $\mathcal{Q}$ *restricts to product distributions:*

$$q(\theta) = q_1(\theta_1)\cdots q_r(\theta_r).$$

7. Another choice for D *could be (low/high) entropy regularization:* $D(q) := \pm H(q)$.

8. Let $p_0(x)$ *be an expected data distribution then we could also use the regularizer (for some divergence* d*):*

$$D(q) := \mathbb{E}_{q(\theta)}[d(p_\theta||p_0)].$$