Take-Home Problem

DO NOT DISTRIBUTE

1 Introduction

In this problem we explore an alternative to learning the parameters of a model by optimising the log-likelihood; namely, we will assume a prior over them and perform approximate inference.

2 Posterior Inference

Suppose we have a model with parameters θ and likelihood $p(\mathcal{D} \,|\, \theta)$. Given data set $\mathcal{D}$, we could choose θ via optimising the log-likelihood:

$$\theta = \operatorname*{argmax}_{\theta'} \log p(\mathcal{D} \,|\, \theta'). \tag{1}$$

There are two problems with this approach: we have a *prior* belief $p(\theta)$ about θ, but this belief is not incorporated in Equation (1); and optimising $p(\mathcal{D} \,|\, \theta)$ with respect to θ gives us a value for θ, but does not tell us how confident we should be in that estimate. To tackle both issues, we can instead find the *posterior* distribution $p(\theta \,|\, \mathcal{D})$, which tells us what we should believe about θ after observing the data $\mathcal{D}$:

$$p(\theta \,|\, \mathcal{D}) = \frac{1}{Z} p(\theta) p(\mathcal{D} \,|\, \theta). \tag{2}$$

Despite its simplicity, Equation (2) is hard to compute: Z requires one to integrate $p(\theta)p(\mathcal{D} \mid \theta)$ over θ, and this integral is often intractible. To compute Equation (2), we must resort to approximate techniques.

3 Stein Variational Gradient Descent

Variational inference[1] is a commonly-used technique to approximate difficult distributions like Equation (2). Specifically, given a family of tractable distributions $\mathcal{Q}$, variational inference seeks to find the one that approximates $p(\theta \mid \mathcal{D})$ best:

$$q(\theta) = \operatorname*{argmin}_{q' \in \mathcal{Q}} \mathrm{D}_{\mathrm{KL}}(q'(\theta) \,\|\, p(\theta \mid \mathcal{D})) \tag{3}$$

where D_{KL} denotes the Kullback-Leibler divergence. A recently-developed technique, called Stein Variational Gradient Descent (SVGD), attempts to perform the minimisation in Equation (3) in the following iterative manner: Denote the distribution resulting from a bijective, differentiable change of variables T in q by $q_{[T]}$, and denote the identity function by id. Letting $T = \mathsf{id}$—note that $q_{[\mathsf{id}]} = q$—SVGD slightly pertubs T in the direction that most decreases $\mathrm{D}_{\mathrm{KL}}(q_{[T]}(\theta) \,\|\, p(\theta \mid \mathcal{D}))$:

$$T \leftarrow \mathsf{id} - \varepsilon \frac{\delta}{\delta T} \mathrm{D}_{\mathrm{KL}}(q_{[T]}(\theta) \,\|\, p(\theta \mid \mathcal{D}))\Big|_{T=\mathsf{id}}. \tag{4}$$

where the functional derivative should be interpreted in the context of some vector-valued reproducing kernel Hilbert space $\mathcal{H}$, equipped with kernel k. Then, for small enough ε,

$$q \leftarrow q_{[T]} \tag{5}$$

should slightly change q and slightly decrease $\mathrm{D}_{\mathrm{KL}}(q(\theta) \,\|\, p(\theta \mid \mathcal{D}))$. SVGD iterates Equation (4) and Equation (5) to solve the minimisation problem in Equation (3), which is most easily implemented in terms of samples of $q(\theta)$: Appendix A shows that a sample $\hat{\theta}$

[1] For an overview of variational inference, please refer to [Mur12]; [Bis06].

from $q(\theta)$ can be transformed to a sample from $q_{[T]}(\theta)$ via

$$\hat{\theta} \leftarrow \hat{\theta} + \varepsilon \mathbb{E}_{q(\theta)}[k(\theta, \hat{\theta}) \nabla_\theta \log p(\theta \mid \mathcal{D}) + \nabla_\theta k(\theta, \hat{\theta})].$$

Here $k(\theta, \hat{\theta}) \nabla_\theta \log p(\theta \mid \mathcal{D})$ pushes a sample $\hat{\theta}$ to high probability regions of p, whilst $\nabla_\theta k(\theta, \hat{\theta})$ pushes the sample θ away from other samples.

4 Problems

P1 Implement SVGD and redo the toy example from [LW16].

P2 Consider the following probabilistic model:

$$\begin{aligned} (a, b, c) := \theta &\sim Prior(\theta), \\ f &= a \cdot x^c + b, \\ \epsilon &\sim N(0, \sigma^2), \\ y &= f + \epsilon \end{aligned}$$

Choose a $Prior(\theta)$ and a sensible value for σ^2; draw a toy data set $\mathcal{D}$ (20–50 data points) from $y \mid \theta$ given a sample $\theta \sim Prior(\theta)$; and compute $p(\theta \mid \mathcal{D})$ using SVGD.

P3 Also estimate θ via Maximum Likelihood (Equation (1)), yielding $\theta^{(\mathrm{MLE})}$. Compare $\theta^{(\mathrm{MLE})}$ to $p(\theta \mid \mathcal{D})$, and compare the prediction of f by the above model ($p(f \mid \mathcal{D})$) to the prediction of f if one were to fix $\theta = \theta^{(\mathrm{MLE})}$ instead ($p(f \mid \mathcal{D}, \theta = \theta^{(\mathrm{MLE})})$).

A Functional Derivative in Equation (4)

First, compute

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\,\mathrm{D}_{\mathrm{KL}}(q_{[\mathsf{id}+\varepsilon\phi]}(\theta)\,\|\,p(\theta\,|\,\mathcal{D}))\Big|_{\varepsilon=0} \\
&\quad= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\,\mathrm{D}_{\mathrm{KL}}(q(\theta)\,\|\,p_{[(\mathsf{id}+\varepsilon\phi)^{-1}]}(\theta\,|\,\mathcal{D}))\Big|_{\varepsilon=0} \\
&\quad= -\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathbb{E}_{q(\theta)}[\log p((\mathsf{id}+\varepsilon\phi)(\theta)\,|\,\mathcal{D}) + \log|\det\nabla_\theta(\mathsf{id}+\varepsilon\phi)(\theta)|]\Big|_{\varepsilon=0}.
\end{aligned}$$

Here,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\log p((\mathsf{id}+\varepsilon\phi)(\theta)\,|\,\mathcal{D})\Big|_{\varepsilon=0} = \phi^{\mathsf{T}}(\theta)\nabla_\theta\log p(\theta\,|\,\mathcal{D}) = \langle\phi, k(\theta,\cdot)\nabla_\theta\log p(\theta\,|\,\mathcal{D})\rangle_{\mathcal{H}},$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\log|\det\nabla_\theta(\mathsf{id}+\varepsilon\phi)(\theta)|\Big|_{\varepsilon=0} = \operatorname{tr}\nabla_\theta\phi(\theta) = \nabla_\theta^{\mathsf{T}}\phi(\theta) = \langle\phi,\nabla_\theta k(\theta,\cdot)\rangle_{\mathcal{H}}.$$

Therefore, plugging in the above equations,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\,\mathrm{D}_{\mathrm{KL}}(q_{[\mathsf{id}+\varepsilon\phi]}(\theta)\,\|\,p(\theta\,|\,\mathcal{D}))\Big|_{\varepsilon=0} &= -\mathbb{E}_{q(\theta)}[\langle\phi, k(\theta,\cdot)\nabla_\theta\log p(\theta\,|\,\mathcal{D})\rangle_{\mathcal{H}} + \langle\phi,\nabla_\theta k(\theta,\cdot)\rangle_{\mathcal{H}}] \\
&= \langle\phi, -\mathbb{E}_{q(\theta)}[k(\theta,\cdot)\nabla_\theta\log p(\theta\,|\,\mathcal{D}) + \phi, \nabla_\theta k(\theta,\cdot)]\rangle_{\mathcal{H}},
\end{aligned}$$

which shows that

$$\frac{\delta}{\delta T}\,\mathrm{D}_{\mathrm{KL}}(q_{[T]}(\theta)\,\|\,p(\theta\,|\,\mathcal{D}))\Big|_{T=\mathsf{id}} = -\mathbb{E}_{q(\theta)}[k(\theta,\cdot)\nabla_\theta\log p(\theta\,|\,\mathcal{D}) + \nabla_\theta k(\theta,\cdot)].$$

References

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006 (cit. on p. 2).

[LW16] Qiang Liu and Dilin Wang. "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm". In: *Advances in Neural Information Processing Systems*. 29. Curran Associates, Inc., 2016, pp. 2378–2386 (cit. on p. 3).

[Mur12] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012 (cit. on p. 2).