Notes on ADVI

Wilker Aziz

October 17, 2017

Consider a generative model of some observed data x

$$p(x, \theta) = p(\theta)p(x|\theta) \tag{1}$$

where θ is a continuous latent random variable possibly constrained to a subset of $\mathbb{R}^d$. For generality, we assume that the posterior distribution

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\int p(\theta')p(x|\theta')\mathrm{d}\theta'} \tag{2}$$

is not available in closed-form due to a generally intractable marginalisation.

ADVI (Kucukelbir et al., 2016) is a black-box inference technique based on variational inference (VI) (Jordan et al., 1999; Blei et al., 2017) and automatic differentiation. ADVI assumes

- a differentiable observation model $p(x|\theta)$;
- a posterior $p(\theta|x)$ whose support is that of the prior $p(\theta)$;
- a bijective transformation from the (possibly constrained) support of the prior to $\mathbb{R}^d$;
- a reparameterisable variational family;

In VI, we maximise a lowerbound on the marginal likelihood (or evidence) of the data (ELBO)

$$\mathcal{E}(\phi|x) = \mathbb{E}_{q_\phi(\theta)}\left[\log \frac{p(x, \theta)}{q_\phi(\theta)}\right] \tag{3}$$

where ϕ indexes a member of a variational family. VI thus turns inference into an optimisation problem which under certain circumstances can be approached by stochastic gradient-based optimisation—ADVI is a an inference technique which satisfies these circumstances for a large class of models. In order to perform unconstrained optimisation with respect to variational parameters, we express the ELBO in terms of a model transformed

with respect to the support of the random variable θ. That is, we employ an invertible transformation mapping from the support of $p(\theta|x)$ to $\mathbb{R}^d$

$$\zeta = t(\theta) \tag{4a}$$
$$\theta = t^{-1}(\zeta) \tag{4b}$$

and optimise an alternative ELBO

$$\mathcal{E}(\phi|x) = \mathbb{E}_{q_\phi(\zeta)}\left[\log \frac{p(x,\zeta)}{q_\phi(\zeta)}\right] = \mathbb{E}_{q_\phi(\zeta)}\left[\log p(x, \theta = t^{-1}(\zeta)) + \log\left|\det J_{t^{-1}(\zeta)}\right|\right] + \mathbb{H}\left[q_\phi(\zeta)\right] \tag{5}$$

where $J_{t^{-1}(\zeta)}$ is the Jacobian matrix of the inverse transformation.[1] Note that we only need a variational family for the unconstrained variable.[2]

Automatic differentiation (Baydin et al., 2015) is a very powerful tool that allows us to approach optimisation of arbitrarily complex functions for as long as they are differentiable. An objective such that of (5) is in general not tractable, but because we choose $q_\phi(\zeta)$ we can make sure that obtaining an MC estimate of the ELBO is trivial. Stochastic optimisation (Robbins and Monro, 1951) is another powerful black-box tool whereby we can attain a local optima of a surface by taking noisy, though unbiased, gradient steps. Unfortunately, the expression in (5) is not ready for stochastic gradient-based optimisation, since the density which we take the expectation with respect to, i.e. $q_\phi(\zeta)$, depends on the parameters we mean to differentiate with respect to, i.e. ϕ. It turns out that for some variational families, which we will call *reparameterisable* (Kingma and Welling, 2014; Rezende et al., 2014), we can express the ELBO in terms of a distribution which is independent of variational parameters. Location-scale distributions are such that we have an invertible transformation that absorbs the parameters of the distribution,[3] i.e.

$$\epsilon = s_\phi(\zeta) = C^{-1}(\zeta - \mu) \tag{6a}$$
$$\zeta = s_\phi^{-1}(\epsilon) = \mu + C\epsilon \tag{6b}$$

where $\phi = \{\mu, C\}$ consists of a location parameter μ and scale parameter C such that $\det C > 0$,[4] and whose Jacobians are

$$J_{s_\phi(\zeta)} = C^{-1} \tag{7a}$$
$$J_{s_\phi^{-1}(\epsilon)} = C \tag{7b}$$

[1] A change of variable applied to a probability density function requires scaling by the absolute determinant of the Jacobian: $p_{X,Y}(x,y) = p_{X,Z}(x, f^{-1}(y))\left|\det J_{f^{-1}(y)}\right|$.

[2] Here we have the first smart move towards black-box inference: transform the model first, so that the variational family already meets the requirements for unconstrained optimisation.

[3] Ruiz et al. (2016) introduce a more general technique to reparameterise densities that are not location-scale.

[4] Note that the invertible transformation depends on variational parameters.

Consider the general problem of estimating $\mathbb{E}_{q(\zeta|\mu,C)}\left[f(\zeta)\right]$ (8a): we first employ integration by substitution (8b) in order to integrate over the support of s_ϕ^{-1}; then we employ a change of density (8c) to express the expectation with respect to parameter-free $q(\epsilon)$.

$$\mathbb{E}_{q(\zeta|\mu,C)}\left[f(\zeta)\right] = \int q(\zeta|\mu,C) f(\zeta)\,\mathrm{d}\zeta \tag{8a}$$

$$= \int q(\zeta = s_\phi^{-1}(\epsilon)|\mu,C) f(\zeta = s_\phi^{-1}(\epsilon)) \underbrace{\left|\det J_{s_\phi^{-1}(\epsilon)}\right| \mathrm{d}\epsilon}_{\text{change of variable}} \tag{8b}$$

$$= \int \underbrace{q(\epsilon = s(\zeta))\left|\det J_{s_\phi(\zeta)}\right|}_{\text{change of density}} f(\zeta = s_\phi^{-1}(\epsilon))\left|\det J_{s_\phi^{-1}(\epsilon)}\right| \mathrm{d}\epsilon \tag{8c}$$

$$= \int q(\epsilon = s_\phi(\zeta))\left|\det C^{-1}\right| f(\zeta = s_\phi^{-1}(\epsilon))|\det C|\,\mathrm{d}\epsilon \tag{8d}$$

$$= \mathbb{E}_{q(\epsilon)}\left[f(\zeta = s_\phi^{-1}(\epsilon))\right] \quad . \tag{8e}$$

We can now express the gradient of the ELBO with respect to variational parameters in terms of $q(\epsilon)$, i.e.

$$\nabla_\phi \mathcal{E}(\phi|x) = \mathbb{E}_{q(\epsilon)}\left[\nabla_\phi \log \frac{p(x, \theta = t^{-1}(\zeta = s_\phi^{-1}(\epsilon)))\left|\det J_{t^{-1}(\zeta = s_\phi^{-1}(\epsilon))}\right|}{q_\phi(\zeta = s_\phi^{-1}(\epsilon))}\right] \tag{9}$$

for which we can trivially obtain an MC estimate since.[5]

References

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2015). Automatic differentiation in machine learning: a survey. *arXiv preprint arXiv:1502.05767*.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted).

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *International Conference on Learning Representations*.

[5] Note that s_ϕ and s_ϕ^{-1} are differentiable functions, thus automatic differentiation will take care of derivatives w.r.t. ϕ via chain rule.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2016). Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China. PMLR.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Ruiz, F. R., AUEB, M. T. R., and Blei, D. (2016). The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*, pages 460–468.