CSCI-GA.2565-001 Machine Learning: Homework 3

Due 11.59 p.m. EST, Dec 19, 2022 on Gradescope

(fill in your name here)
(collaborators if any)

We encourage LaTeX-typeset submissions but will accept quality scans of hand-written pages.

1 Variational Inference and Monte Carlo Gradients

In this question, we will review the details of variational inference (VI), in particular, we will implement the gradient estimators that make VI tractable.

We consider the latent variable model $p(\mathbf{z}, \mathbf{x}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{z}_i)p(\mathbf{z}_i)$ where $\mathbf{x}_i, \mathbf{z}_i \in \mathbb{R}^D$. Recall that in VI, we find an approximation $q_\lambda(\mathbf{z})$ to $p(\mathbf{z}|\mathbf{x})$.

(A) Let $V_1(\lambda)$ be the set of variational approximations $\{q_\lambda : q_\lambda(\mathbf{z}) = \prod_{i=1}^{N} q(\mathbf{z}_i; \lambda_i)\}$ where λ_i are parameters learned for each datapoint $\mathbf{x}_i$. Now consider $f_\lambda(\mathbf{x})$ as a deep neural network with *fixed* architecture where λ parametrizes the network. Let $V_2(\lambda) = \{q_\lambda : q_\lambda(\mathbf{z}) = \prod_{i}^{N} q(\mathbf{z}_i; f_\lambda(\mathbf{x}_i))\}$. Which of the two families (V_1 or V_2) is more expressive, i.e. approximates a larger set of distributions? **Prove** your answer.

Will your answer change if we let f_λ represent *variable* architecture, e.g. if λ parametrizes the set of multi-layered perceptrons of all sizes? Why or why not?

Solution. Write your solution for each question using the solution environment. Feel free to use style packages to your convenience, e.g. highlighting parts of your solution that you still need to work on. □

(B) For variational inference to work, we need to compute unbiased estimates of the gradient of the ELBO. In class, we learnt two such estimators: score function (REINFORCE) and pathwise (reparametrization) gradients. Let us see this in practice for a simpler inference problem.

Consider the dataset of $N = 100$ one-dimensional data points $\{x_i\}_{i=1}^{N}$ in `data.csv`. Suppose we want to minimize the following expectation with respect to a parameter μ:

$$\min_{\mu} \mathop{\mathbb{E}}_{z\sim\mathcal{N}(\mu,1)} \left[\sum_{i=1}^{N} (x_i - z)^2 \right] \qquad (1)$$

(i) Write down the score function gradient for this problem. Using a suitable reparametrization, write down the reparameterization gradient for this problem.

(ii) Using PyTorch and for each of these two gradient estimators, perform gradient descent using $M = \{1, 10, 100, 1000\}$ gradient samples for $T = 10$ trials. Plot the mean and variance of the final estimate for μ for each value of M across the T trials.

You should have two graphs, one for each gradient estimator. Each of the graph should contain two plots, one for the means and one for the variances. The x-axis should be M, hence each of these plots will have four points.

(C) What conditions do you require on $p(z)$ and $f(z)$ ($f(z) = \sum_{i=1}^{N}(x_i - z)^2$ in this case) for each of the two gradient estimators to be valid? Do these apply to both continuous and discrete distributions $p(z)$?

2 Bayesian Parameters versus Latent Variables

(A) Consider the model $y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$ where the inverse-variance is distributed $\lambda = 1/\sigma^2 \sim \text{Gamma}(\alpha, \beta)$. Show that the predictive distribution $y^\star | \mathbf{w}, \mathbf{x}^\star, \alpha, \beta$ for a datapoint $\mathbf{x}^\star$ follows a generalized T distribution

$$T(t; \nu, \mu, \theta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\theta\sqrt{\pi\nu}} \left(1 + \frac{1}{\nu}\left(\frac{t-\mu}{\theta}\right)^2\right)^{-\frac{\nu+1}{2}}$$

with degree $\nu = 2\alpha$, mean $\mu = \mathbf{w}^\top \mathbf{x}^\star$ and scale $\theta = \sqrt{\beta/\alpha}$. You may use the property $\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$.

(B) Using your expression in (A), write down the MLE objective for $\mathbf{w}$ on N arbitrary labelled datapoints $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Do not optimize this objective.

(C) Now consider the model $y_i \sim \mathcal{N}\Big(f(\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}), \sigma^2\Big)$ where $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, σ^2 is known, and f is a deep neural network parametrized by $\mathbf{w}$.

(i) Write down an expression for the predictive distribution $y^\star | \mathbf{X}, \mathbf{y}, \mathbf{x}^\star$, where $\mathbf{X}, \mathbf{y}$ denote the training datapoints. *(You may leave your answer as an integral.)*

(ii) Describe how you would approximate this distribution using variational inference and how you can use your approximation to make a prediction for $\mathbf{x}^\star$. Your answer should include the distribution $p(\cdot)$ that you wish to approximate *(which may or may not be the predictive distribution itself)*, the distribution $q(\cdot)$ that is the variational approximation, as well as the variational objective.

(D) Finally, consider the model $y \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2)$ where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, I)$ and σ^2 is known.

Derive a closed-form expression for the predictive distribution $y^\star | \mathbf{X}, \mathbf{y}, \mathbf{x}^\star$. What are the parameters of this predictive distribution and how do you optimize them?

(E) Of the three models defined in parts (A), (C), and (D) above, which are latent variable models and which are not? Why? *(If any are ambiguous, explain why.)*

(F) Of the three models defined in parts (A), (C), and (D) above, which are Bayesian models and which are not? Why? *(If any are ambiguous, explain what is Bayesian about it and what is not.)*

3 Normalizing Flows

In this question, we will review how we can use invertible transformations and the change-of-variables formula to turn simple distributions into complex ones. Such transformations are known as *normalizing flows*. One reason that flows are useful is that they can map unimodal distributions into multimodal ones, while still allowing for a tractable density.

(A) Let $z_0 \sim \mathcal{N}(0,1)$. Produce a density plot of z_0 using $N = 1000$ samples.

Now look up "Planar Flow" (Equation 4) of The Expressive Power of a Class of Normalizing Flow Models. Denote this flow as f. Choose an invertible non-linearity h and find values of w, b, u such that $f(z_0)$ is a multimodal distribution. Plot the density plot of $f(z_0)$ using the same $N = 1000$ samples as above.

Note that $d = 1$ in this question. Also, it is fine to choose a h that is only invertible on its output range, e.g. the sigmoid function on $(0, 1)$.

(B) Use the change-of-variables formula and write down an explicit expression for the density of $f(z_0)$. This depends on your choice of h.

(C) Let's generalize this to D-dimensional variables and to a sequence of invertible functions f (not necessarily planar flows). We can sample $\mathbf{z}^{(0)} \sim \mathcal{N}(\mathbf{0}, I)$ and then transform that sample iteratively using a sequence of invertible functions, $f_1, \ldots, f_K$, to finally obtain a sample $\mathbf{z}^{(K)}$ where

$$\mathbf{z}^{(K)} = f_K \circ \cdots \circ f_2 \circ f_1(\mathbf{z}^{(0)}).$$

Using the change-of-variables formula, write down a formula for the log-density of $\mathbf{z}^{(K)}$ using the functions $\{f_k\}_{k=1}^{K}$, their inverses and Jacobians, as well as the log-density of $\mathbf{z}^{(0)}$.

(D) Let us consider how we can improve variational inference using flows.

(i) How can we define an approximation $q_\lambda(\mathbf{z})$ to $p(\mathbf{z}|\mathbf{x})$ using flows?

(ii) How do we train the model, i.e. optimize λ to yield a good approximation to $p(\mathbf{z}|\mathbf{x})$?

(iii) In what way is this more flexible than using a Gaussian approximation for $q_\lambda(\mathbf{z})$?

4 Causal Inference: Doubly Robust Estimators

Denote unit i's treatment, outcome, and its vector of covariates by T_i, Y_i, and X_i. Let us model propensity scores by $e(x;\hat{\theta})$ and the outcome by $f(x;\hat{\psi})$. Assume strong ignorability and positivity hold. We define the Doubly Robust Estimator (DRE) for the average outcome when treated, $\mathbb{E}[Y^{(1)}]$, by:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i}{e(X_i;\hat{\theta})} - \frac{T_i - e(X_i;\hat{\theta})}{e(X_i;\hat{\theta})} f(X_i;\hat{\psi})\right] \tag{2}$$

(A) Suppose the propensity model is correctly specified, i.e. $e(x;\hat{\theta}) = P(T_i = 1|X_i = x)$. Given any function f, show that the DRE is unbiased.

(B) Suppose the model $f(x;\hat{\psi})$ is correctly specified, i.e. $f(x;\hat{\psi}) = \mathbb{E}[Y_i^{(1)}|X_i = x] := \mathbb{E}[Y_i|X_i = x, T = 1]$. Given any function e taking values in $(0,1)$, show that the DRE is unbiased.

(C) Recall that control variates improve Monte Carlo estimators by defining a new estimator with the same expectation but lower variance. For some estimator f, this is done by taking a function g with $\mathbb{E}[g(x)] = 0$, and defining the new estimator as $\hat{f}(x) = f(x) - ag(x)$ for some $a \in \mathbb{R}$. Here $\mathbb{E}[\hat{f}] = \mathbb{E}[f]$, but the variances are not equal. The value of a that makes the variance of $\hat{f}$ smallest is $a^* = \frac{Cov(f,g)}{Var(g)}$.

When both $e(x;\hat{\theta})$ and $f(x;\hat{\psi})$ are correctly specified, use control variates to justify the use of Doubly Robust Estimators.

5 Generative Models with f-divergences

Given two distributions P and Q with density functions p and q, we define the f-divergence as:

$$D_f(P\|Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

with respect to a convex function f.

(A) How can we estimate f-divergences using likelihood-ratio estimators as we learned in class?

(B) Using the estimate from your solution to (A), show how we can minimize $D_f(P\|Q_\theta)$ with respect to θ.

(C) Let $\mathcal{Q}$ denote the family of distributions that Q_θ lives in. Assume that $P \notin \mathcal{Q}$. Compare the KL and Reverse KL divergences. What are properties that each f-divergence imposes on its corresponding minimizer $Q_\theta^\star$, i.e.

$$Q_\theta^\star = \arg\min_{Q_\theta \in \mathcal{Q}} D_f(P\|Q_\theta)$$

Name	$D_f(P\|Q)$	$f(u)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} dx$	$u \log u$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} dx$	$-\log u$

6 Reinforcement Learning with Sparse Rewards

Suppose that you have a robot that should learn to move from any of a set of starting positions $s_0 \in \mathcal{S}_0$ to a goal position G. The robot can move by performing a sequence of small, low-level continuous actions, such as rotating its joints. However, you do not know how to explicitly program a sequence of actions that will move the robot from any s_0 to G. You decide to use a reinforcement learning algorithm. Your RL algorithm uses the policy gradient to learn a policy $\pi_\theta(a|s)$ that maps from states s to distributions over actions a. You consider learning episodes of a finite number of T steps.

(A) You start by designing the Markov Decision Process (MDP) that defines the robot's learning environment. The robot receives reward $R = 1000$ when it reaches location G and $R = 0$ upon entering any other position. What is the Monte Carlo estimate of the policy gradient when G is not reached in T steps in any of the sample trajectories? What happens as the robot explores in this environment and what will its learning process look like?

(B) Suppose the robot must start in s_0 for all learning trajectories. Name two ways you could alter the MDP environment to improve exploration and gradients.