

DD2434/FDD3434 Machine Learning, Advanced Course

Assignment 1AD, 2025

Harald Melin, Jens Lagergren

Deadline, see Canvas

Read this before starting

There are some commonalities between the problems and they cover different aspects of the course and vary in difficulty, consequently, it may be useful to read all of them before starting. Also think about the formulation and try to visualize the model. You are allowed to discuss the formulations, but have to make a note of the people you have discussed with. You will present the assignment by a written report and code, submitted before the deadline using Canvas. Furthermore, there will be an oral exam after the deadline where you have to show understanding of your solutions in order to keep your passing score. You may use AI tools to assist you in the writing (See Canvas for use of AI in the course), but you must ensure you understand any solution you provide. You must solve the assignment individually and it will automatically be checked for similarities to other students' solutions as well as documents on the web in general. Although you are allowed to discuss the problem formulations with others, you are not allowed to discuss solutions.

From the report it should be clear what you have done and you need to support your claims with results. You are supposed to write down the answers to the specific questions detailed for each task. This report should clearly show how you have drawn your conclusions and explain your derivations. Your assumptions, if any, should be stated clearly. Show the results of your experiments using images and graphs together with your analysis and add your code as an appendix.

Being able to communicate results and conclusions is a key aspect of scientific as well as corporate activities. It is up to you as an author to make sure that the report clearly shows what you have done. Based on this, and only this, we will decide if you pass the task. No detective work should be required on our side. In particular, neat and tidy reports please!

Grading system

This assignment is divided into grade levels D, C, B and A. In order to receive a particular grade, you must pass the criteria for that level and all levels below it. Each subproblem contains one or more questions. Each subproblem is graded Pass/Fail. Passing criteria for each level:

- D** Passing 2/3 subproblems.
- C** Passing 3/4 subproblems.
- B** To come.
- A** To come.

These grades are valid for assignments submitted before the deadline, late assignments can at most receive the grade E, which makes it meaningless to hand in late solutions for this assignment.

Good Luck!

I D - level

For the D-level, there are two theory questions, 1.D.1 and 1.D.2, and two practice/implementation questions 1.D.3 and 1.D.4. Choose only ONE out of 1.D.3 and 1.D.4 to answer. Note that 1.C.3 depends on 1.D.3.

1.1 Theory - D level

1.D.1

Variational inference approximates the true posterior $p(Z | X)$ using a tractable variational distribution $q(Z)$. Consider a latent-variable model with joint density $p(X, Z)$, posterior $p(Z | X)$, and an arbitrary variational distribution $q(Z)$. Answer the following queries.

Question 1.1.1: Starting from the definition of the KL divergence $\text{KL}(q(Z) \| p(Z | X))$, show that it can be rewritten in the form

$$\log p(X) = \mathcal{L}(q) + \text{KL}(q(Z) \| p(Z | X)),$$

and identify the quantity referred to as the Evidence Lower Bound (ELBO).

Question 1.1.2: In one sentence each, describe how the choice of variational family (e.g., a fully factorized mean-field distribution versus a more expressive structured distribution) affects

1. the tightness of the ELBO, and
2. the accuracy of the posterior approximation.

1.D.2

Assume that we have a mean field assumption where our variational distribution q factorizes over three different variables, such that

$$q(Z_1, Z_2, Z_3) = q_1(Z_1)q_2(Z_2)q_3(Z_3) \quad (1)$$

Let the joint distribution be denoted as $p(X, Z)$, $Z = \{Z_1, Z_2, Z_3\}$.

Denote by q_1^* the q_1 that maximizes the ELBO.

Question 1.1.3: Prove that q_1^* satisfies $\log q_1^*(Z_1) = \mathbb{E}_{-Z_1}[\log p(X, Z)]$.

Hint: You may take inspiration from Jens' video lecture about CAVI.

1.2 Practice/implementation - D level

1.D.3

Consider the model with Normal-likelihood and NormalGamma prior of 1E.3. In this exercise, you should apply Black-Box VI with the REINFORCE estimator instead of CAVI to infer $q(\mu)$ and $q(\tau)$. Use the same mean-field assumption of 1E.3, i.e., $q(\mu, \tau) = q(\mu)q(\tau)$.

Question 1.2.4: *Provide expressions for the log-likelihood, log-prior, log-variational distributions and score functions of the variational distributions. No derivations are needed, only final expressions.*

Question 1.2.5: *Implement Algorithm 1 of the BBVI paper [Ranganath et al., 2014]. You should use Pytorch or numpy for the implementation, either using automatic differentiation or computing gradients explicitly. Reuse the data sampling script of 1E.3. and prior parameter values, but provide results only for the case $N = 100$. Provide two plots in the report: 1. showing the ELBO over iterations, 2. the Expected values $\mathbb{E}_q[\mu]$ and $\mathbb{E}_q[\tau]$.*

1.D.4

In this exercise we revisit the “GMM-light” model from Exercise Session 3, but now treat the component precisions (τ_k) as unknown. As before, for $n = 1, \dots, N$ and $k = 1, \dots, K$ we have

$$p(x_n | z_n = k, \mu_k, \tau_k) = \mathcal{N}(x_n | \mu_k, \tau_k^{-1}), \quad p(z_n | \pi) = \text{Categorical}(z_n | \pi), \quad (2)$$

where π is assumed known.

We now place a conjugate Normal–Gamma prior on (μ_k, τ_k) :

$$p(\mu_k, \tau_k) = \text{NormalGamma}(\mu_k, \tau_k | m_0, \lambda_0, a_0, b_0), \quad (3)$$

Question 1.2.6: *Draw the directed graphical model / Bayes net for this model.*

Question 1.2.7: *Write down the complete-data joint density*

$$p(x_{1:N}, z_{1:N}, \mu_{1:K}, \tau_{1:K})$$

up to proportionality, and then its logarithm $\log p(x_{1:N}, z_{1:N}, \mu_{1:K}, \tau_{1:K})$.

Question 1.2.8: *State the mean-field variational family you will use.*

Question 1.2.9: *Using the generic CAVI update*

$$\log q_j^*(\theta_j) = \mathbb{E}_{-j} [\log p(x, \theta)] + \text{const},$$

derive the optimal variational factor $q^(\mu_k)$ for each component k .*

II C - level

For the C-level, there are two theory questions, 1.C.1 and 1.C.2, and two practice/implementation questions 1.C.3 and 1.C.4.

2.1 Theory - C level

1.C.1

Instead of our regular ELBO, [Burda et al., 2015] proposed the Importance-Weighted ELBO, or IWELBO for short, defined as:

$$\mathcal{L}_K := \mathbb{E}_{Z_1, \dots, Z_K} \left[\frac{1}{K} \sum_{k=1}^K \log \frac{p(X_k, Z_k)}{q(Z_k | X_k)} \right] \quad (4)$$

where $K \in \mathbb{N}$ is the number of samples.

Question 2.1.10: Show that the IWELBO is a valid lower bound on the log-marginal likelihood, i.e. show that $\mathcal{L}_K \leq \log p(X)$.

Question 2.1.11: Show that the IWELBO tightens the variational bound for K samples, i.e. that $\mathcal{L}_K \geq \mathcal{L}_1$ for $K > 1$, where \mathcal{L}_1 denotes the regular ELBO given by

$$\mathcal{L}_1 := \mathbb{E}_Z \left[\log \frac{p(X, Z)}{q(Z | X)} \right].$$

1.C.2

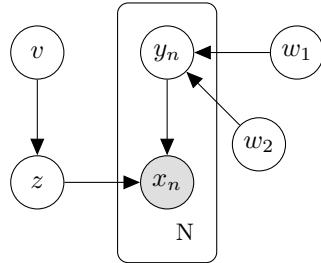


Figure 1: 1.C.2 PGM of some generic model for Rao-Blackwellization.

Consider the model described in figure 1, and the mean-field approximation:

$$q(w_1, w_2, z, v, y) = q_{\lambda_1}(w_1)q_{\lambda_2}(w_2)q_{\lambda_3}(z)q_{\lambda_4}(v) \prod_n q_{\lambda_{5,n}}(y_n). \quad (5)$$

Question 2.1.12: Derive the Rao-Blackwellized partial gradient of the ELBO w.r.t. λ_3 , $\nabla_{\lambda_3} \mathcal{L}$ following the steps of [Ranganath et al., 2014]. Write out the final expression for the Rao-Blackwellized $\nabla_{\lambda_3} \mathcal{L}$.

2.2 Practice/implementation - C level

1.C.3

Question 2.2.13: Extend the implementation of problem 1.D.3 with the Control variate used in the BBVI paper [[Ranganath et al., 2014](#)], i.e., implement Algorithm 2 but without Rao-Blackwellization. Provide the same plots as for 1.D.3.

1.C.4

In this question, we study the Gamma distribution and use its exponential-family structure to derive the Fisher Information Matrix (FIM) and implement Natural Gradient Descent (NGD). We consider the Gamma distribution parameterized by shape $\alpha > 0$ and scale $\beta > 0$:

$$p(x | \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right), \quad x > 0. \quad (6)$$

Question 2.2.14: Rewrite the Gamma distribution in canonical exponential-family form,

$$p(x | \eta) = h(x) \exp(\eta^\top t(x) - A(\eta)),$$

and identify the natural parameters η , the sufficient statistics $t(x)$, and the log-normalizer $A(\eta)$. Derive the gradient $\nabla_\eta A(\eta)$.

Question 2.2.15: Consider an arbitrary differentiable loss function $L(\theta)$ for this model, where $\theta = (\alpha, \beta)$. Using the properties of exponential-family distributions and the Fisher information, express the natural gradient $\tilde{\nabla}_\theta \mathcal{L}$ as a function of the standard gradient $\nabla_\theta \mathcal{L}$ and the FIM. (You do not need to specify the explicit entries of the FIM in this sub-question.)

Question 2.2.16: Compute the inverse Fisher Information Matrix $F(\alpha, \beta)^{-1}$ explicitly. You may verify your expression numerically.

Question 2.2.17: Let $x_{1:N}$ be i.i.d. samples from $\text{Gamma}(\alpha^*, \beta^*)$. Write down the average negative log-likelihood

$$\mathcal{L}(\alpha, \beta) = -\frac{1}{N} \sum_{n=1}^N \log p(x_n | \alpha, \beta),$$

and derive its gradient with respect to (α, β) .

Question 2.2.18: Implement a Python notebook that:

- Samples $N = 1000$ points from $\text{Gamma}(\alpha^* = 3.0, \beta^* = 2.0)$.
- Initializes (α, β) to a poor guess.
- Estimates (α, β) using **standard gradient descent (GD)**.
- Estimates (α, β) using **natural gradient descent (NGD)**, using the definition

$$\tilde{\nabla}_{\theta}\mathcal{L} = F(\theta)^{-1}\nabla_{\theta}\mathcal{L}, \quad \theta = (\alpha, \beta).$$

- Ensures $\alpha > 0$ and $\beta > 0$ during optimization.

Question 2.2.19: Plot and compare the convergence trajectories of GD and NGD for both parameters (α_t, β_t) , as well as the evolution of the negative log-likelihood. In one or two sentences, describe why the natural gradient $\tilde{\nabla}_{\theta}\mathcal{L}$ leads to faster and more stable convergence.

References

- [Burda et al., 2015] Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.
- [Ranganath et al., 2014] Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR.