“Amortized Variational Inference: Towards the Mathematical Foundation and Review” Sept. 2022

Authors: Ankush Ganguly, Sanjana Jain, Ukrit Watchareeruetai
Link to actual paper: https://arxiv.org/abs/2209.10888

1. The key idea of variational inference is to convert the statistical inference problem of computing the posterior probability density into a tractable optimization problem.
2. This property enables variational inference to be faster than several sampling-based approaches.
3. The traditional variational inference algorithm is unable to scale to large data sets…
4. … and is unable to readily infer out-of-bounds data points without re-running the optimization process.
5. Generative modeling tasks make use of amortized variational inference for its efficiency and scalability…
6. … as it uses a parameterized function to learn the parameters of the approximate posterior.
7. Amortized variational inference issues include:...
8. … generalization issues, the amortization gap, …
9. … inconsistent representation learning, and posterior collapse.
10. Approximate inference provides solutions to non-conjugate models for which analytic posteriors are unavailable.
11. Conjugacy occurs where the posterior is in the same family of probability density functions as the prior, …
12. … but with new parameter values which have been updated to reflect learning from the data.
13. Variational inference deals with inefficient approximate inference by the use of a suitable metric to select the best tractable approximation to the true posterior.
14. Variational inference takes advantage of the speed benefits of maximum a posteriori (MAP) estimation.
15. Other optimization-based inference techniques include loopy-belief propagation and expectation maximization.
16. The traditional 1999 VI algorithm introduces a new set of parameters, characterizing the approximate density, for every new observation.
17. This leads to inefficient scalability as the number of parameters grows linearly with the number of observations.
18. Amortized inference uses a stochastic function to estimate the true posterior.
19. The parameters of this stochastic function are fixed and shared across all data points, thereby amortizing the inference.
20. Deep neural networks are a popular choice for this stochastic function as…

21. … they combine probabilistic modeling with the representational power of deep learning.
22. Amortized inference combined with deep neural networks has been shown to scale efficiently to large data sets.
23. The VAE and its variants are examples of this.
24. We assume the observed data points are independently and identically distributed, …
25. … and are generated by some random process involving the unknown random variables.
26. We assume that for each observed data point, there is a latent variable with some prior probability density.
27. We assume the data points are sampled from the conditional probability distribution, which is also the generative model.
28. In our case, we can think of the data points as images, …
29. … and the latent variables as low-dimensional representations of those images.
30. From a coding theory perspective, the latent variables can be seen as code and thus form the basis for representation learning.
31. We use Bayes theorem to compute the posterior probability distribution as: p(z|x) = p(x|z)p(z) / p(x) where p(x) = integral of p(x|z)p(z) with respect to z.
32. The evidence, aka the marginal likelihood, for most statistical models is high dimensional
33. We compute the evidence to evaluate a chosen model's ability to fit the data.
34. The traditional VI algorithm is as follows: for each observed data point, x_i, select an approximate posterior from a family of tractable densities, Q.
35. Each approximate density is characterized by a set of their own variational parameters…
36. … and is a candidate approximation to the true posterior evaluated at data point x_i.
37. The goal is to tune these parameters to get an optimal approximation to the true posterior.
38. The complexity and accuracy of the optimization process depends on the choice of the variational family, …
39. … which depends on a "measure" that captures the difference between the approximate and the true posterior.
40. Usually, this "measure" is chosen to be the non-negative KL divergence.
41. The optimization problem for traditional VI is to reduce the relative entropy…
42. … by choosing the approximate density with the lowest reverse KL divergence to the true posterior, sampling one data point at a time.
43. The output of this optimization is a set of variational parameters that characterize the best approximation to the true posterior.
44. Thus, for each local variational parameter, inference amounts to solving the following optimization problem:
45. q*(z|x_i; ξ_i) = argmin over q \in Q of KL(q(z|x_i; ξ_i) || p(z|x_i; theta))
46. In the case of VI, the forward KL divergence cannot be computed in closed form…
47. … as it requires taking expectations with respect to the unknown posterior.
48. Su et al. 2018 proved that GANs, like VAEs, are a special case of VI…
49. … and proposed a unified framework between the two by reformulating the VI objective.

50. VI enables efficient computation of a lower bound to the evidence.
51. A better fit to the observed data by a statistical model requires a better fit to the evidence by that model.
52. Sometimes the evidence lower bound (ELBO) is used as a basis for selecting models to fit the data distribution.
53. In traditional VI, the ELBO is the sum of the negative reverse KL divergence and the log evidence.
54. Another way to compute the ELBO is to partition the latent variable z into k disjoint groups z_k where k is a natural number between 1 and N.
55. This factorized form of the VI objective corresponds to a framework developed in physics known as mean field theory, and is known as mean field VI.
56. An extension to mean field VI is structured VI, which adds dependencies between the variables leading to a better approximation of the true posterior.
57. The coordinate ascent algorithm can look like the EM algorithm…
58. … where the "E step" computes approximate conditionals of local latent variables, …
59. … and where the "M step" computes an approximate conditional of the global latent variables.
60. Similar to mean field VI, this optimization process works by repeatedly updating the variational paramaters of each random variable…
61. … based on the variational parameters of the random variables in its Markov Blanket,
62. … and re-estimating the convergence of the ELBO.
63. The Markov Blanket of a target variable is a minimal set of variables that the target variable is conditioned on, …
64. … while all other remaining variables in the model are probabilistically independent of the target variable.
65. Stochastic variational inference combines natural gradients and stochastic optimization to solve the scalability issue of the traditional VI algorithm.
66. Coordinate ascent variational inference updates the variational parameters one data point at a time.
67. Stochastic variational inference uses stochastic optimization on a subsample of the data…
68. … and updates the variational parameters based on that sub-sample.
69. The methodology of SVI is to get a stochastic estimator of the ELBO…
70. … based on a set of M samples at each iteration with or without replacement.
71. This allows us to take derivates, …
72. … and update the local variational parameters based on the M samples…
73. … as well as the global variational parameter, theta, using stochastic gradient ascent.
74. We repeat this process until the ELBO converges.
75. A Euclidean gradient points in the direction of steepest ascent in a Euclidean space.
76. A natural gradient points in the direction of steepest ascent in a Riemannian space, …

77. … a space where local distance is defined via the symmetric KL-divergence rather the L2 norm.
78. do Carmo in 1993 introduced a Riemannian metric, $I(\xi)$, which defines the distance between ξ and a nearby vector $\Delta\xi$ as $\Delta\xi$^T * $I(\xi)$ * $\Delta\xi$ as approximately equal to the…
79. … the symmetric KL divergence between ξ and $\xi+\Delta\xi$, …
80. … where $I(\xi)$ is the Fisher information matrix of $q(z; \xi)$.
81. The Fisher information matrix is essential to compute the Cramer-Rao lower bound…
82. …. for the performance analysis of an unbiased estimator, …
83. … a minimum variance estimator for a parameter.
84. In VI, for a high dimensional parameter space, …
85. … studying the covariance matrix for the variational estimator provides an estimate for its unbiasedness.
86. The underlying high dimensional posterior structure might be rich, …
87. … and the covariance matrix for the variational parameters captures…
88. … the uncertainty of the KL divergence being locked onto one of its many local modes.
89. Additionally, the covariance matrix for the variational parameters captures the sensitivity of the estimated posterior density with respect to small variations in the variational parameters.
90. For the variational parameters to be unbiased estimators of the true parameters…
91. … they must satisfy the Cramer-Rao lower bound as…
92. … $cov(\xi) \geq [I(\xi)]$^-1
93. Additionally, the Fisher information matrix is a measure of the curvature for a probability density function…
94. … as it is equal to the expected Hessian for that probability density function.
95. This property is useful where the Fisher information matrix is infeasible to store, invert, or convert.
96. In such cases, computing the second moment of the derivates is equivalent to approximating the Fisher information matrix.
97. The speed of convergence for the SVI optimization process depends on the variance of the gradient estimates.
98. Lower variance in the gradient estimates ensures minimum gradient noise, …
99. … allowing for larger learning rates which leads to faster convergence.
100. One way to reduce the variance of the gradient estimates is to increase the mini-batch size, …
101. … which leads to lower gradient noise as suggested by the law of large numbers.
102. Another approach to reduce the variance of the gradient estimates is to to use non-uniform sampling, such as importance sampling, …
103. … to select mini-batches with lower gradient noise.
104. Hardware memory constraints might make increasing the mini-batch size implausible.

105. Another approach to increase the speed of the training procedure is to adjust the learning rate while keeping the mini-batch size fixed.
106. The idea is to let the empirical gradient variance guide the adaptation of the learning rate, …
107. … which is inversely proportional to the gradient noise at each iteration.
108. Gradually adjusting the learning rate guarantees that every point in the parameter space can be reached, …
109. … while the gradient noise decreases sufficiently fast to ensure convergence.
110. Another approach to reduce the variance is to use a control variate, …
111. … a stochastic term, which when added to the stochastic gradient, …
112. … reduces the variance while keeping its expected value intact.
113. Using control variates for variance reduction is common in Monte Carlo simulation and stochastic optimization.
114. The traditional VI process requires an initial, analytical derivation of the ELBO, which requires time and mathematical expertise.
115. Ranganath et al. 2014 introduced the BBVI methodology that removes the need for the analytical computation of the ELBO, …
116. … expanding applications beyond conditionally conjugate exponential families.
117. Note that the score function and sampling algorithms depend only on the variational distribution, not the underlying model.
118. BBVI enables the practitioner to obtain an unbiased gradient estimator by sampling without having to derive the the ELBO explicitly.
119. But the variance of the gradient estimates under the Monte Carlo estimates can be too large to be useful.
120. In stochastic variational inference, subsampling from a finite set of data points leads to high noise in the gradient estimates.
121. However, in BBVI, it is the possible oversamping of the random variables that leads to high noise in the gradient estimates.
122. Variance reduction techniques for BBVI include:
123. … the combination of Rao-Blackwellization and control variates, …
124. … local expectation gradients, …
125. … overdispersed importance sampling, …
126. … and the reparameterization trick.
127. Both Kingma and Welling (2013) and Ranganath et al. (2014) state that the Monte Carlo gradients in BBVI exhibit high variance.
128. So, Kingma and Welling (2013) introduced a more practical gradient estimator in the form of a reparameterization trick.
129. For a chosen approximate posterior, the trick allows a random variable to be a differentiable transformation of a noise variable.

130. After applying the trick to the ELBO for VI, we get the stochastic estimator for the ELBO.
131. In the stochastic estimator, the gradient of the log joint distribution is included as part of the expectation.
132. The advantage of including the gradient of the log joint distribution in the expectation is that this term is more informed about the direction of the maximum posterior mode.
133. This information also attributes to the lower variance for the gradient estimates when compared to the policy gradient estimates.
134. The reparameterization trick is the basis of VAEs.
135. A property of general purpose inference algorithms is that they are memoryless, …
136. … where each observation is processed independently of the others.
137. So, inference using one observation will not interfere with inference using another observation.
138. There is no mechanism to resume the knowledge from previous inferences on newer ones.
139. Inferring on the same observation twice and two separate observations requires the same amount of computation.
140. To keep a memory trace of past inferences, although at a higher cost, to solve the scalability issue of traditional VI.
141. Amortizing the inference means flexible memoized reuse of past inferences to compute inferences on newer observations.
142. To this end, amortized VI makes use of a stochastic function, which maps the observed variable to the latent variable belonging to the variational posterior, …
143. … the parameters of which are learned during the optimization process.
144. Instead of having separate parameters for each observation, …
145. … the estimated function can infer latent variables even for new data points without rerunning the optimization process all over again on the new data points.
146. In traditional VI, a local variational parameter is introduced for every observation
147. In amortized VI, the variational parameters are shared globally across observations.
148. VAEs employ two deep neural networks: a probabilistic encoder and a probabilistic decoder.
149. A probabilistic decoder is a top-down generative model that creates a mapping from a latent variable z_i to a data point x_i aka a generative network
150. A probabilistic encoder is a bottom-up inference model that approximates the posterior probability density aka a recognition network.
151. In amortized VI, the ELBO is the sum of the expected log likelihood and the KL divergence…
152. … between the approximate density and the prior over the latent variable evaluated at individual data points.
153. The KL divergence term can be interpreted as regularizing phi, …

154. … encouraging the approximate posterior to be close to the prior.
155. There are two connections here to auto-encoders: (1) the KL divergence term acts as a regularizer, and (2) the expected log likelihood is the expected negative reconstruction error.
156. To get a tighter ELBO and hence better variational approximations, importance sampling can be used to get a lower variance estimate of the evidence.
157. The approximation gap can be reduced by choosing a variational family that is flexible enough to contain the true posterior as one solution.
158. The concept of normalizing flow… to improve the expressiveness of the variational approximation.
159. A normalizing flow describes the transformation of a probability density function through a sequence of invertible mappings.
160. It involves repeatedly applying change of variables to transform the simple initial approximation into a richer approximation to better match the true posterior.
161. The idea of auxiliary variables has been employed in hierarchical variational models,
162. … where dependencies between latent variables are induced similarly to the induction of dependencies between data in hierarchical Bayesian models.
163. Amortizing the inference introduces a coding efficiency gap known as the amortization gap.
164. The complexity of the variational density determines the approximation gap.
165. The capacity of the stochastic function determines the amortization gap.
166. The amortization gap and the approximation gap contribute to the inference gap, …
167. … which is the gap between the marginal log likelihood and the log ELBO.
168. Shu et al. 2018 introduces amortized inference regularization that restricts the capacity of the encoder, …
169. … to prevent both the inference and the generative models from overfitting to the training set.
170. Vanilla VAEs are not auto-encoding, i.e., …
171. … samples from the generative network are not mapped to the corresponding representations by the recognition network.
172. The optimal denoising VAE model is a kernel regression model, …
173. … and the variance of the injected noise controls the smoothness of the optimal recognition model.
174. Posterior collapse occurs when the variational posterior lies close or collapses to the prior.
175. This causes the generative network to ignore a subset of the latent variables.
176. Hence, the model fails to learn a valuable representation of the data.
177. The zero-forcing nature of the reverse KL divergence helps to concentrate on one mode rather than spread mass over all of them.
178. Zero-forcing leads to underestimating of the posterior variance.

179. It leads to degenerate solutions during optimization and is the source of pruning in VAEs.
180. The KL-divergence is a special case of a family of divergence measures known as the alpha-divergences.
181. Choosing different alpha values allows the variational approximation to balance between…
182. … zero-forcing (alpha approaching infinity) and mass-covering (alpha approaching negative infinity) behavior.
183. Alpha divergences are a subset of a more general family of divergences known as f-divergences.
184. What are the alternatives to the non-convex ELBO?
185. The variational Hölder bound is a convex upper bound to the evidence, …
186. … the minimization of which is a convex optimization problem that can be solved using existing convex optimization algorithms.
187. Generally, the distance between points in the latent space in a VAE…
188. … does not reflect the true similarity of corresponding points in the observation space.
189. To improve the representation learning in VAEs
190. To understand the geometrical properties of the latent space in VAEs
191. VAEs lack the ability to take into account the uncertainty in posterior approximation in a principled manner.
192. To make posterior approximation in VAEs more interpretable by using Bayesian Neural Networks…
193. … as the choice for the parametric functions for both the inference and the generative models in VAEs.