Problem Set: Mathematics of Generative Modelling

April 29, 2023

0.1 Optimization

Exercise 0.1

Short answer questions.

1. Define convex set and convex function.
2. Give example of a function that is not convex.
3. Define conjugate of a function.
4. Show that the conjugate function is always convex.
5. Define strong and weak duality.
6. What are the optimization solvers used for min-max problems in GAN or WGAN?
7. What is Jensen's inequality.
8. State log-sum inequality and prove it using Jensen's inequality.
9. Define lower semi-continuous function.
10. Is lower semi-continuous function necessarily convex? Give example.

Exercise 0.2

Long Answer Questions.

1. Write an optimization problem with equality and inequality constraints.
2. Write the dual problem.
3. Show that the dual function is concave.
4. Derive a min-max characterization for weak and strong duality (for inequality constraints only).
5. What is Slater's condition?
6. If f^* denote the conjugate of f, then when do we have $f^{**} = f$?

0.2 Probability and Statistics

Exercise 0.3

Answer the following.

1. Define independent and mutually exclusive events.
2. Define conditional expectation.
3. Write PDF of Bernoulli, Binomial, Standard normal and Normal distribution.
4. When do we say that two random variables are independent?
5. Define marginal PDF for a function of two random variables.
6. Define law of total variance, law of total expectation, law of total probability.
7. Define correlation and covariance.
8. State method of transform for vector of random variables.
9. State Markov's, Chebychev's, Cauchy-Schwartz, Jensen's inequality.
10. Define convergence in probability, convergence in mean, convergence in distribution.
11. State law of large numbers.
12. State central limit theorem (CLT).
13. Explain how CLT is used in practice in estimations.
14. Define sample mean.
15. What is MLE.
16. What is MAP estimation.
17. Sample mean an unbiased estimator for true mean.

18. Define a push-forward distribution.
19. How can we generate samples from $N(\mu, \sigma^2)$ from $N(0, 1)$? Here N stands for normal distribution.
20. State how inverse CDF can be used for generating samples given the initial samples $x_1, \dots, x_m$.
21. More concretely, let us assume that we did an experiment to obtain following samples:

$$S = \{1, 1, 3, 1, 5, 2, 4, 2, 7, 1, 2, 3, 5, 6, 6, 3, 9, 9, 1, 3, 3, 4, 8\}.$$

We know that these samples are generated from the set $\{1, \dots, 9\}$. However, we don't know the underlying distribution. Hence we don't know the real distribution. That is, we don't know the probability of each numbers. Use inverse CDF method to generate more samples for this underlying distribution.

22. What are the difficulties for generating samples using inverse CDF approach?
23. Prove the relationship between MSE, Bias and Variance.
24. Prove that if $MSE(\hat{\Theta})_n$ goes to zero and n tends to ∞, then $\hat{\Theta}_n$ is an consistent estimator.
25. A bag contains 3 balls. Each ball is either red or blue. Let θ be the number of blue balls. Here $\theta = 0, 1, 2, 3$. Define the samples $(X_1, X_2, X_3, X_4) = (1, 1, 0, 1)$. Estimate number of blue balls.

0.3 Probability Measures

Exercise 0.4

1. Define KL divergence.
2. Show that $KL(p, q)$ is convex in p and q.
3. Show an example to show that $KL(p, q) \neq KL(q, p)$.
4. Define $f-$divergence.
5. Derive $KL-$divergence as a special case of $f-$divergence.
6. Define JS divergence.
7. Define Wassertein divergence.
8. Define total variation divergence.
9. Give example where KL divergence can become infinite but Wasserstein distance may not.
10. What is smoothing in the context of KL divergence?
11. What could be negative effects of smoothing?
12. Show that Minimizing KL divergence is equivalent to MLE.
13. Name the paper where $f-$divergence was proposed.

0.4 GAN and WGAN

Exercise 0.5

Answer the following with True or False.

1. Wassertein distance is a metric.
2. Wassertein distance satisfies triangle inequality.
3. If K Lipschitz functions are allowed, then wassertein distance scales by K factor.
4. Both GAN and WGAN explicitely construct the probability density.

Exercise 0.6

Answer the following with short answers.

1. Wassertein distance can be derived from a general optimal transport problem.
2. Wassertein distance is a metric.

3. Wassertein distance satisfies triangle inequality.
4. What is the generic name of parametrized functions taken in GAN or WGAN?
5. Write the Wassertein GAN loss function (min-max) form.
6. What is the role of clipping parameter in WGAN.
7. The number of iterations for discriminator is usually less than the number of iterations for generator weight updates.
8. WGAN distance induces a strong topology whereas JS imposes a strong topology.
9. In the calculation for expectation in the WGAN and GAN algorithm, we do average sum over samples. Justify why this is done.
10. Name the paper where gradient penalty was introduced.

Exercise 0.7

Answer the following long answer questions. Give derivation if required.

1. Derive GAN loss function from KL divergence.
2. Motivate why discriminator is seen as a classifier by looking at the final GAN loss.
3. What is mode collapse?
4. Write full algorithm for GAN training.
5. After training GAN, how are the samples generated?
6. For a given generator G, derive the optimal discriminator.
7. Derive that optimal generator cost is $-\log 4$.
8. Derive that the best cost for generator is achieved when $p_g = p_{data}$.
9. Derive the WGAN loss function by proving KR-duality.
10. Write WGAN algorithm.
11. Write the advantages of WGAN over GAN.
12. State the proposition that suggests why adding gradient penalty can improve WGAN. No need to prove this.

Exercise 0.8

Answer the following long answer questions. Give derivation if required.

1. Define an optimal transport problem. As a special case, describe Wasserstein distance.
2. Describe entropy regularized optimal transport. What is the role of entropy regularization? https://youtu.be/kPXMDTYzxEA?t=292
3. Show that linear function plus strongly convex function is strongly convex. Let $f(x)$ be linear and $g(x)$ be strongly convex function.
4. Derive alternating projection algorithm to solve entropy regularized OT. https://youtu.be/dgQb8_KBDFU?t=2110
5. Write the Sinkhorn algorithm. https://youtu.be/kPXMDTYzxEA?t=1518
6. Define a matrix scaling problem. https://youtu.be/kPXMDTYzxEA?t=1564
7. Prove that the primal iterates (with alternating projection method) is equivalent to Sinkhorn algorithm. https://youtu.be/kPXMDTYzxEA?t=2955
8. Write algorithm for Sinkhorn generative modelling. https://youtu.be/kPXMDTYzxEA?t=3852
9. Draw architecture model to describe Sinkhorn generative modelling. https://youtu.be/kPXMDTYzxEA?t=4162
10. Define in one line probabilistic graphical model. https://youtu.be/aqPCEOKj2as?t=342
11. Show two examples of probabilistic graphical models and joint probabilities. https://youtu.be/aqPCEOKj2as?t=680
12. Why VAE is called a latent variable model?

13. Derive evidence lower bound for VAE, write VAE objective with loss used, repramaterization trick used, and training procedure. Also, show schematic diagram of VAE.
14. Describe hierarchical VAE and Markovian HVAE. https://youtu.be/oQcgmMMJPPI
15. Write names of three papers related to Variational Diffusion Models. https://youtu.be/oQcgmMMJPPI?t=1852
16. Describe modelling assumptions for variational diffusion models. https://youtu.be/oQcgmMMJPPI?t=2210
17. What are the implications of the assumptions above? https://youtu.be/oQcgmMMJPPI?t=2639
18. Write down the expressions for $p(x_{0:T})$ and $p(x_T)$ for Variational diffusion model. https://youtu.be/oQcgmMMJPPI?t=3362
19. Describe in words to layman in 5 lines how diffusion models work. https://youtu.be/oQcgmMMJPPI
20. Derive evidence lower bound to obtain reconstruction term, prior matching term, and consistency term. (starts https://youtu.be/oQcgmMMJPPI?t=4116 and ends in L16: 6:39).
21. Derive that $q(x_{t-1}|x_t, x_0)$ is a normal distribution. https://youtu.be/-XuX6scRrE8?t=2479
22. Derive that minimizing $D_{KL}(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t))$ is equivalent to minimizing $\|\hat{x}_\theta(x_t, t) - x_0\|_2^2$. https://youtu.be/X4F8WFA0u88?t=638
23. Show how diffusion noise parameters are learned using signal to noise ratio. https://youtu.be/X4F8WFA0u88?t=2148
24. Derive a VDM formulation where the noise ϵ_0 is predicted. https://youtu.be/X4F8WFA0u88?t=3692
25. Derive VDM formulation where score is matched. https://youtu.be/X4F8WFA0u88?t=4966
26. Show how learning to model a score is equivalent to modeling the negative of the source noise. https://youtu.be/X4F8WFA0u88?t=6090
27. State and prove implicit score matching. https://youtu.be/-tk3JmLQPOk?t=2930
28. Show how conditional diffusion models are used for both classifier and classifier free guidance. https://youtu.be/F5z-m9koRRc?t=218
29. Show 2 differences between autoregressive flows and normalizing flows. https://youtu.be/F5z-m9koRRc?t=2965
30. Define first order autoregressive process. https://youtu.be/F5z-m9koRRc?t=4182
31. Define wide sense stationary process. https://youtu.be/F5z-m9koRRc?t=5526
32. Describe fully visible sigmoid belief network and Neural Autoregressive Distribution Estimation. https://youtu.be/Y70irj3_rjU?t=916
33. Describe RNN and LSTM networks. https://youtu.be/gEAK3toYUbM?t=75
34. What is the advantage of LSTM over RNN?
35. Describe seq2seq model with attention. https://youtu.be/Ys7A01eEf90?t=402
36. Describe sub-word tokenization.
37. Describe Transformer architecture for an input $X \in \mathbf{R}^{t \times d}$: explain scaled self attention, show diagram of multihead attention with 3 heads, explain positional encoding, layer normalization and skip connection. Then explain decoder block: masked self attention, cross multihead attention. https://youtu.be/SNf7AeXal9o?t=1845 https://youtu.be/e_ddKtAI2Ao