CS 480/680 Homework 6
Released March 15, 2026
NOT GRADED
©Marina Meilă
mmp@uwaterloo.ca

All plots must be included in the .pdf writeup

Problem 1 – SGD and Heavy Ball
1.a Repeat Problem 2 from Assignment 5, part **g** by replacing Gradient Descent (GD) with Stochastic Gradient Descent (SGD) with $n' = 1, 10, 100, 1000$. Compare the number of epochs Tn'/n to convergence and compare also with GD from your previous homework.

1.b Use now the heavy ball method with all the methods above and observe the change in learning curves.

Problem 2 – KL divergence
The KL divergence between two distributions q, p on $\mathbb{R}^m$ is given by $KL(q||p) = \int_{\mathbb{R}^m} \ln \frac{q(z)}{p(z)} q(z) dz$.
2.a Let $m = 1$, i.e $z \in \mathbb{R}$. Derive the formula for $KL(\mathcal{N}(\mu, \sigma^2)||\mathcal{N}(0,1))$. Bring the expression to its simplest form for full credit.

2.b Now let $z \in \mathbb{R}^m$ with $m > 1$, and let $\mu \in \mathbb{R}^m$, $\Sigma = \text{diag}\{\sigma_j^2,\ j = 1:m\}$ (diagonal covariance matrix).

Fact 1 $z_{1:m}$ are mutually independent.

Fact 2 If $p(z_1, z_2) = p_1(z_1)p_2(z_2)$ and $q(z_1, z_2) = q_1(z_1)q_2(z_2)$, then $KL(q||p) = KL(q_1||p_1) + KL(q_2||p_2)$.

Use Facts 1 and 2 to derive the formula for $KL(\mathcal{N}(\mu, \Sigma)||\mathcal{N}(0, I_m))$. Bring the expression to its simplest form for full credit.

Extra exercise Prove Facts 1 and 2.

Problem 3 – ELBO
In VAE, $\mu_\phi(x), s_\phi(x), \pi_\theta(z)$ are three neural networks that output respectively $\mathbb{E}_\phi[(Z|x)] \in \mathbb{R}^m$, $\ln \sigma_j(Z|x)$ for $j = 1:m$, and $Prob[X_k = 1|z]$ for $k = 1:d$. Let the following denote the gradients of these networks w.r.t. the parameters and inputs

$$g_{\mu,\phi} = \frac{\partial \mu_\phi(x)}{\partial \phi} \qquad g_{s,\phi} = \frac{\partial s_\phi(x)}{\partial \phi} \qquad g_{\pi,\theta} = \frac{\partial \pi_\theta(x)}{\partial \theta} \tag{1}$$

$$g_{\mu,x} = \frac{\partial \mu_\phi(x)}{\partial x} \qquad g_{s,x} = \frac{\partial s_\phi(x)}{\partial x} \qquad g_{\pi,z} = \frac{\partial \pi_\theta(z)}{\partial z} \tag{2}$$

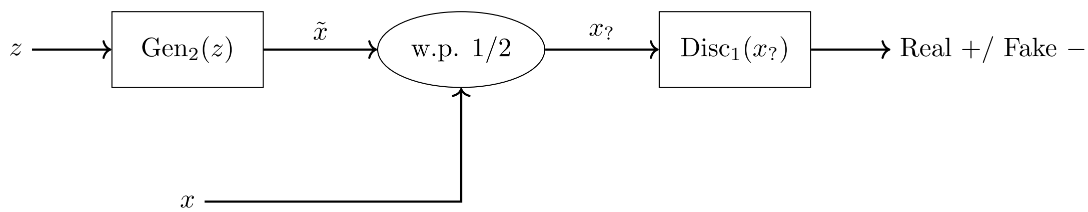
NOTE in the above p_θ was now changed to π_θ to agree with the disambiguated notation in Lecture VI.

3.a Derive the gradient w.r.t. ϕ of the expression (14) in Lecture VI (the KL divergence term of the ELBO) as a function of the gradients above.

3.b To evaluate expression (15) (first term of the ELBO) n_ϵ samples from $\mathcal{N}(0, I_m)$ are taken. Derive the gradient w.r.t. ϕ of the r.h.s. of expression (15) in Lecture VI as a function of the gradients above.

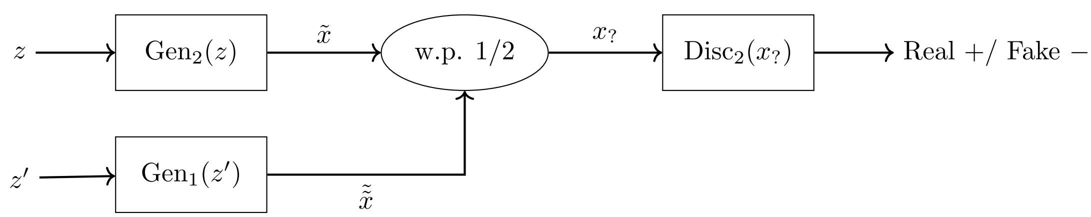
3.c Under the same conditions as above, derive the gradient w.r.t. θ of the r.h.s. of expression (15) in Lecture VI as a function of the gradients above.

Problem 4 – Generative Adversarial Networks
Assume that data come from a (true) distribution q.
1.a You have successfully trained $(\text{Gen}_1, \text{Disc}_1)$ on q and now Gen_1 generates data with distribution $p_1 \approx q$. Your colleague researcher Gooz wants to train a new encoder Gen_2 on the same q. Gen_2 has a different architecture that Gooz believes is superior. Gooz wants to speed up training of Gen_2 by using the "pre-trained" Disc_1 with the untrained Gen_2.



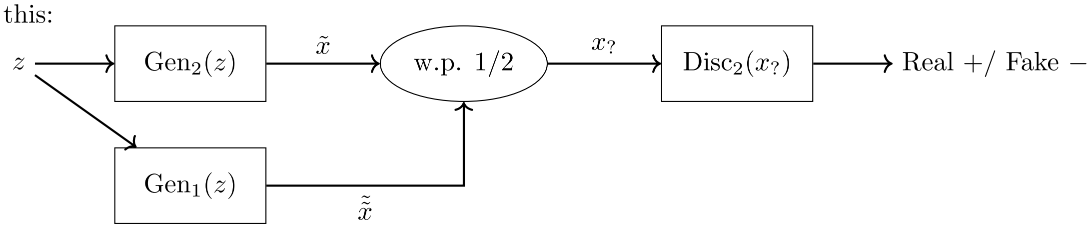
Will this architecture accelerate the training of Gen_2 to reconstruct q? Explain why or why not.

1.b A week later, Gooz decides to train their $(\text{Gen}_2, \text{Disc}_2)$ from scratch. Unfortunately they misplaced the hard drive containing the training data. They decide to use the "pre-trained" Gen_1 to generate a new training data like this:



Will this architecture successfully train Gen_2 to reconstruct q? Explain why or why not.

1.c Gooz reconsiders the architecture in **1.b** and decides to use a single random generator, like this:



Will this architecture successfully train Gen_2 to reconstruct q? Explain why or why not.