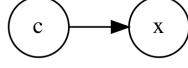


# 1 InfoGAN

InfoGAN extends the GAN objective to include a new term which encourages high mutual information between generated data and a subset of latent codes [1]. Let  $(\mathbf{c}, \mathbf{z})$  be latent variable, where  $\mathbf{c}$  are latent codes capturing semantic features of the data distribution and  $\mathbf{z}$  are source of incompressible noise.

## 1.1 Probabilistic Interpretation



A simpler view of method presented in the paper is to consider the above generative model. The joint density can be factorized as follows

$$p_{\mathbf{c}, \mathbf{x}} = p_{\mathbf{c}}(c)p_{\mathbf{x}|\mathbf{c}}(x|c) = \prod_{l=1}^L p_{c_l}(c_l)p_{\mathbf{x}|\mathbf{c}}(x|c)$$

The paper implicitly model  $p_{\mathbf{x}|\mathbf{c}}$  by using a combination of 1) a deterministic generator  $G : \mathcal{C} \times \mathcal{Z} \rightarrow \mathcal{X}$  and 2) a stochastic noise sampler  $\mathbf{z} \sim p_{\mathbf{z}}$ . In particular,  $f : \mathcal{C} \rightarrow \mathcal{X}; c \mapsto G(c, z)$  for some  $z \sim p_{\mathbf{z}}$  is trained to sample from  $p_{\mathbf{x}|\mathbf{c}}(\cdot|c)$  using the adversarial loss [2].

## 1.2 Variational Maximization of Mutual Information

The paper is motivated to construct latent code in such a way such that when given a sample, we would be quite certain what the latent codes are. In other words, we are interested in the following optimization problem

$$\min_G H(\mathbf{c}|\mathbf{x}) \quad \text{where} \quad \mathbf{x} = G(\mathbf{c}, \mathbf{z}) \quad (1)$$

If we know the parametric family of distribution  $\mathbf{c}$  is in, this is equivalent to maximizing mutual information between latent codes and generated sample. Given  $H(\mathbf{c}|\mathbf{x}) = H(\mathbf{c}) - I(\mathbf{c}; \mathbf{x})$ , we can rewrite (1) as

$$\max_G I(\mathbf{c}; \mathbf{x}) = \mathbb{E}_{\mathbf{c}, \mathbf{x}} \left[ \ln \frac{p_{\mathbf{c}, \mathbf{x}}(c, x)}{p_{\mathbf{c}}(c)p_{\mathbf{x}}(x)} \right]$$

which is intractable, since we do not know the implicit likelihood  $p_{\mathbf{x}|\mathbf{c}}$  nor the posterior  $p_{\mathbf{c}|\mathbf{x}}$ . Instead we approximate  $p_{\mathbf{c}|\mathbf{x}}$  with using  $q_{\mathbf{c}|\mathbf{x}}$ , parameterize by a neural network, and derive a lower bound for the objective [3, 4],

$$\begin{aligned} I(\mathbf{c}; \mathbf{x}) &= H(\mathbf{c}) - H(\mathbf{c}|\mathbf{x}) \\ &= \sum_x p_{\mathbf{x}}(x) \sum_c p_{\mathbf{c}|\mathbf{x}}(c|x) \ln p_{\mathbf{c}|\mathbf{x}}(c|x) + H(\mathbf{c}) \\ &= \sum_x p_{\mathbf{x}}(x) \sum_c p_{\mathbf{c}|\mathbf{x}}(c|x) \ln \frac{p_{\mathbf{c}|\mathbf{x}}(c|x)}{q_{\mathbf{c}|\mathbf{x}}(c|x)} + \sum_x p_{\mathbf{x}}(x) \sum_c p_{\mathbf{c}|\mathbf{x}}(c|x) \ln q_{\mathbf{c}|\mathbf{x}}(c|x) + H(\mathbf{c}) \\ &= \mathbb{E}_{\mathbf{x}} [KL(p_{\mathbf{c}|\mathbf{x}}(c|x) || q_{\mathbf{c}|\mathbf{x}}(c|x))] + \mathbb{E}_{\mathbf{c}, \mathbf{x}} [\ln q_{\mathbf{c}|\mathbf{x}}(c|x)] + H(\mathbf{c}) \\ &\geq \mathbb{E}_{\mathbf{c}, \mathbf{x}} [\ln q_{\mathbf{c}|\mathbf{x}}(c|x)] + H(\mathbf{c}) \quad (KL \geq 0) \end{aligned}$$

This lower bound can be optimized using stochastic gradient via Monte Carlo estimation,

$$\begin{aligned}
\nabla_{\theta} I(\mathbf{c}; \mathbf{x}) &= \nabla_{\theta} \mathbb{E}_{\mathbf{c}, \mathbf{x}} [\ln q_{\mathbf{c}|\mathbf{x}}(\mathbf{c}|\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{c}, \mathbf{z}} [\nabla_{\theta} \ln q_{\mathbf{c}|\mathbf{x}}(\mathbf{c}|G(\mathbf{c}, \mathbf{z}))] \\
&\approx \sum_{i=1}^N \nabla_{\theta} \ln q_{\mathbf{c}|\mathbf{x}}(\mathbf{c}^{(i)}|G(\mathbf{c}^{(i)}, \mathbf{z}^{(i)})) \\
&\quad \text{where } \mathbf{c}^{(i)} \sim p_{\mathbf{c}} \quad \mathbf{z}^{(i)} \sim p_{\mathbf{z}} \quad i = 1, \dots, N
\end{aligned}$$

We could also interpret the idea of randomizing the generator using a noise sampler as performing the reparameterization trick [5]. We avoid taking gradient of expectation with respect to  $p_{\mathbf{x}|\mathbf{c}}$ ; Instead, we take sample from a known distribution  $\mathbf{z} \sim p_{\mathbf{z}}$  and then compute the desired sample  $\mathbf{x} = G(\mathbf{c}, \mathbf{z})$  via a deterministic function.

## References

- [1] Xi Chen et al. “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets”. In: *arXiv:1606.03657 [cs, stat]* (June 11, 2016). arXiv: [1606.03657](https://arxiv.org/abs/1606.03657). URL: <http://arxiv.org/abs/1606.03657> (visited on 12/04/2019).
- [2] Ian J. Goodfellow et al. “Generative Adversarial Networks”. In: *arXiv:1406.2661 [cs, stat]* (June 10, 2014). arXiv: [1406.2661](https://arxiv.org/abs/1406.2661). URL: <http://arxiv.org/abs/1406.2661> (visited on 12/12/2019).
- [3] David Barber and Felix Agakov. “The IM Algorithm: A Variational Approach to Information Maximization.” In: Jan. 1, 2003.
- [4] Ben Poole and Sherjil Ozair. “On Variational Bounds of Mutual Information”. In: (2019), p. 10.
- [5] Diederik P. Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *arXiv:1312.6114 [cs, stat]* (May 1, 2014). arXiv: [1312.6114](https://arxiv.org/abs/1312.6114). URL: <http://arxiv.org/abs/1312.6114> (visited on 11/13/2019).