

# Paper Review: Breaking the log-K Curse on Contrastive Learners With FlatNCE

Beomsu Kim\*

July 20, 2021

## Paper Information.

- Junya Chen et. al. Breaking the log-K Curse on Contrastive Learners With FlatNCE. arXiv preprint arXiv:2107.01152, 2021.

## 1 Introduction

- There are many unresolved issues with contrastive learning.
  - Contrastive learners need a very large number of negative samples to work well.
  - The bias, variance, and performance tradeoffs are in debate.
  - There is a lack of training diagnostic tools for contrastive learning.
- Our development starts with two simple intuitions.
  - The contrasts between positive and negative data should be as large as possible.
  - The objective should be properly normalized to yield minimal variance.

---

\*Department of Mathematical Sciences, KAIST. Email `beomsu.kim@kaist.ac.kr`

## 2 Contrastive Representation Learning with InfoNCE

- With  $y_{1:K} = (y_1, \dots, y_K)$ , define

$$p_{XY^K}(x, y_{1:K}) = p_{XY}(x, y_1) \prod_{k \neq 1} p_Y(y_k).$$

This means  $(x, y_1) \sim p_{XY}$  and  $(x, y_k) \sim p_X \otimes p_Y$  for  $k \neq 1$ .

- Let  $g(x, y)$  be a parametrized function, such as a neural network.
- Given  $(x, y_{1:K}) \sim p_{XY^K}$ , define  $g_k = g(x, y_k)$  for  $k = 1, \dots, K$ . Also, define

$$I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g) = \log \frac{\exp(g_1)}{\frac{1}{K} \sum_{k=1}^K \exp(g_k)},$$

and

$$I_{\text{InfoNCE}}^K(X; Y \mid g) = \mathbb{E}_{p_{XY^K}} [I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g)]$$

and

$$I_{\text{InfoNCE}}^K(X; Y) = \max_g I_{\text{InfoNCE}}^K(X; Y \mid g).$$

Note that  $g_1$  is the logit for the “positive pair” and  $g_k$  for  $k \neq 1$  are the logits for the “negative pairs”.

- We also define

$$I_{\text{InfoNCE}}(x, y_{1:K} \mid g) = -\log \frac{\exp(g_1)}{\sum_{k=1}^K \exp(g_k)}$$

such that

$$I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g) = -I_{\text{InfoNCE}}(x, y_{1:K} \mid g) + \log K.$$

**Proposition 2.1.**  $I_{\text{InfoNCE}}^K(X; Y)$  is an asymptotically tight lower bound to mutual information, i.e.,

$$I(X; Y) \geq I_{\text{InfoNCE}}^K(X; Y \mid g), \quad \lim_{K \rightarrow \infty} I_{\text{InfoNCE}}^K(X; Y) \rightarrow I(X; Y).$$

*Proof.* See my paper review on InfoNCE. □

### 3 FlatNCE and Generalized Contrastive Representation Learning

- Define

$$I_{\text{FlatNCE}}(x, y_{1:K} \mid g) = \frac{\sum_{k \neq 1} \exp(g_k - g_1)}{\text{stop\_grad}[\sum_{k \neq 1} \exp(g_k - g_1)]}$$

and

$$I_{\text{FlatNCE}}^K(x, y_{1:K} \mid g) = -\log \frac{1}{K} \sum_{k \neq 1} \exp(g_k - g_1).$$

- We observe that

$$\sum_{k \neq 1} \exp(g_k - g_1) = \left( \frac{\exp(g_1)}{\sum_{k \neq 1} \exp(g_k)} \right)^{-1}$$

and so (the gradient is w.r.t. the parameters of  $g$ )

$$\begin{aligned} \nabla I_{\text{FlatNCE}}(x, y_{1:K} \mid g) &= \frac{\nabla \sum_{k \neq 1} \exp(g_k - g_1)}{\text{stop\_grad}[\sum_{k \neq 1} \exp(g_k - g_1)]} \\ &= \nabla \log \sum_{k \neq 1} \exp(g_k - g_1) \\ &= \nabla \log \frac{1}{K} \sum_{k \neq 1} \exp(g_k - g_1) \\ &= -\nabla I_{\text{FlatNCE}}^K(x, y_{1:N} \mid g). \end{aligned}$$

This shows that gradient descent on  $I_{\text{FlatNCE}}$  is equivalent to gradient descent on  $I_{\text{FlatNCE}}^{\oplus, K}$ .

- We also observe that

$$\begin{aligned} I_{\text{FlatNCE}}^K(x, y_{1:N} \mid g) &= -\log \frac{1}{K} \sum_{k \neq 1} \exp(g_k - g_1) \\ &= -\log \frac{\frac{1}{K} \sum_{k \neq 1} \exp(g_k)}{\exp(g_1)} \\ &= \log \frac{\exp(g_1)}{\frac{1}{K} \sum_{k \neq 1} \exp(g_k)} \end{aligned}$$

which is just  $I_{\text{InfoNCE}}(x, y_{1:K} \mid g)$  with the positive pair logit  $g_1$  removed from the denominator sum.

- Define

$$I_{\text{FlatNCE}}^{\oplus}(x, y_{1:K} \mid g) = \frac{\sum_{k=1}^K \exp(g_k - g_1)}{\text{stop\_grad}[\sum_{k=1}^K \exp(g_k - g_1)]}$$

and

$$I_{\text{FlatNCE}}^{\oplus,K}(x, y_{1:K} \mid g) = -\log \frac{1}{K} \sum_{k=1}^K \exp(g_k - g_1).$$

- Similar to  $I_{\text{FlatNCE}}$ , we have

$$\nabla I_{\text{FlatNCE}}^{\oplus}(x, y_{1:K} \mid g) = -\nabla I_{\text{FlatNCE}}^{\oplus,K}(x, y_{1:K} \mid g)$$

and also

$$I_{\text{FlatNCE}}^{\oplus,K}(x, y_{1:K} \mid g) = \log \frac{\exp(g_1)}{\frac{1}{K} \sum_{k=1}^K \exp(g_k)} = I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g).$$

which shows that gradient ascent on  $I_{\text{FlatNCE}}^{\oplus}$  is equivalent to gradient descent on  $I_{\text{InfoNCE}}^K$ .

**Proposition 3.1.**  $\nabla I_{\text{FlatNCE}}^{\oplus}(x, y_{1:K} \mid g) = \nabla I_{\text{InfoNCE}}(x, y_{1:K} \mid g)$ .

*Proof.* Since (see Section 2)

$$I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g) = -I_{\text{InfoNCE}}(x, y_{1:K} \mid g) + \log K,$$

we have (see the above bullet)

$$\nabla I_{\text{FlatNCE}}^{\oplus}(x, y_{1:K} \mid g) = -\nabla I_{\text{FlatNCE}}^{\oplus,K}(x, y_{1:K} \mid g) = -\nabla I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g) = \nabla I_{\text{InfoNCE}}(x, y_{1:K} \mid g).$$

This concludes the proof.  $\square$

**Proposition 3.2.** *The gradient of  $I_{\text{FlatNCE}}(x, y_{1:K} \mid g)$  is an importance-weighted estimator of the form*

$$\nabla I_{\text{FlatNCE}}(x, y_{1:K} \mid g) = \sum_{k \neq 1} w_k \nabla g_k - \nabla g_1, \quad w_k = \frac{\exp(g_k)}{\sum_{k' \neq 1} \exp(g_{k'})}.$$

*Proof.* Observe that

$$\begin{aligned} \nabla I_{\text{FlatNCE}}(x, y_{1:K} \mid g) &= \frac{\nabla \sum_{k \neq 1} \exp(g_k - g_1)}{\sum_{k \neq 1} \exp(g_k - g_1)} \\ &= \sum_{k \neq 1} \frac{\exp(g_k - g_1)}{\sum_{k' \neq 1} \exp(g_{k'} - g_1)} (\nabla g_k - \nabla g_1) \\ &= \sum_{k \neq 1} \frac{\exp(g_k)}{\sum_{k' \neq 1} \exp(g_{k'})} (\nabla g_k - \nabla g_1) \\ &= \sum_{k \neq 1} w_k (\nabla g_k - \nabla g_1) \\ &= \sum_{k \neq 1} \nabla g_k - \nabla g_1 \end{aligned}$$

since  $\sum_{k \neq 1} w_k = 1$ .  $\square$

- In FlatNCE, larger weights will be assigned to the more challenging negative examples in the batch.
- The authors claim FlatNCE is also a formal MI lower bound using the below Lemma.

**Lemma 3.3.** *For arbitrary  $u \in \mathbb{R}$ , we have*

$$I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g) \geq 1 - u - \frac{1}{K} \sum_{k=1}^K \exp(-u + g_k - g_1)$$

and the inequality holds when

$$u = \text{stop\_grad} \left[ \log \frac{1}{K} \sum_{k=1}^K \exp(g_k - g_1) \right] = \text{stop\_grad} \left[ -I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g) \right].$$

*Proof.* The Fenchel-Legendre dual for  $f(t) = -\log t$  is  $f^*(v) = -1 - \log(-v)$ . That is,

$$f(t) = \sup_{v \in \mathbb{R}} \{vt - f^*(v)\}$$

and so

$$-\log t \geq vt + 1 + \log(-v)$$

for any  $v \in \mathbb{R}$ . Setting  $v = -e^{-u}$ , we get

$$-\log t \geq 1 - u - e^{-u}t$$

for any  $u \in \mathbb{R}$ . Since

$$I_{\text{FlatNCE}}^K(x, y_{1:K} \mid g) = -\log \frac{1}{K} \sum_{k \neq 1} \exp(g_k - g_1),$$

setting

$$t = \frac{1}{K} \sum_{k \neq 1} \exp(g_k - g_1)$$

proves the first part of the proposition. The second part can be checked by simple calculation.  $\square$

**Corollary 3.4.**  $\mathbb{E}_{p_{XYN}} [I_{\text{FlatNCE}}^K(x, y_{1:K} \mid g)] \leq I(X; Y)$ .

*Proof.* Plugging in the optimal value of  $u$  in the above Lemma, we have

$$I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g) = 1 + \text{stop\_grad} [I_{\text{InfoNCE}}^K(x, y_{1:K} \mid g)] - I_{\text{FlatNCE}}^\oplus(x, y_{1:K} \mid g).$$

Since the first two terms at the RHS are constant w.r.t. the parameters of  $g$ , the authors claim that the claimed inequality holds up to a constant.  $\square$

- In my opinion, Lemma 3.3 and Corollary 3.4. contribute essentially nothing to the paper, since it is about  $I_{\text{FlatNCE}}^\oplus$  (which has the same gradients as  $I_{\text{InfoNCE}}^K$  and so suffers from the same gradient vanishing problem).
- If the authors wanted to make a valid contribution, they should have provided analyses about  $I_{\text{FlatNCE}}$ , not  $I_{\text{FlatNCE}}^\oplus$ . The authors claim in the footnote that we can similarly show  $I_{\text{FlatNCE}}$  lower bounds mutual information, but they do not provide any proofs.