Controllable Text Generation

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, Eric P. Xing
School of Computer Science, Carnegie Mellon University
{zhitingh,zichaoy,xiaodan1,rsalakhu,epxing}@cs.cmu.edu

読む人：Akihiko WATANABE
2017/04/24

Paper: https://arxiv.org/pdf/1703.00955.pdf

Controllable Text Generation

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, Eric P. Xing
School of Computer Science, Carnegie Mellon University
{zhitingh,zichaoy,xiaodan1,rsalakhu,epxing}@cs.cmu.edu

読む人：Akihiko WATANABE
2017/04/24

Paper: https://arxiv.org/pdf/1703.00955.pdf

Characteristic of this work:

1. Enable to control attribute of Generated Text

the acting is bad the acting is good	e.g. attribute -> sentiment

i thought the movie was good i guess the movie is good i guess the movie will have been good	e.g. attribute -> tense

2. Propose learning technique to controllable generation

Variational Auto Encoder (VAE) + attribute discriminator

Task: Text Generation

Preliminary Knowledge

How to generate text

(using Neural Network)?

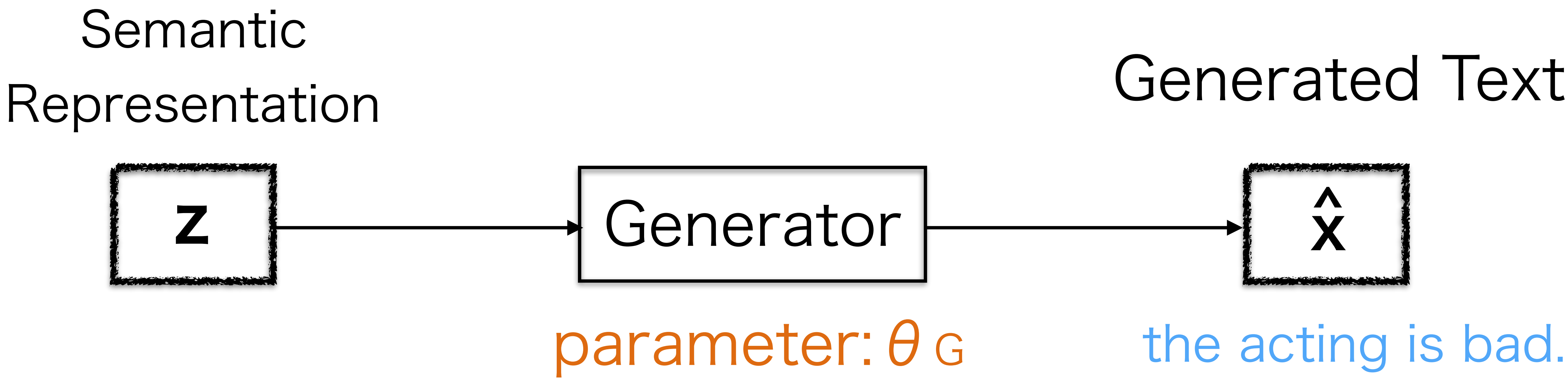
Auto Encoder

※ Caution:

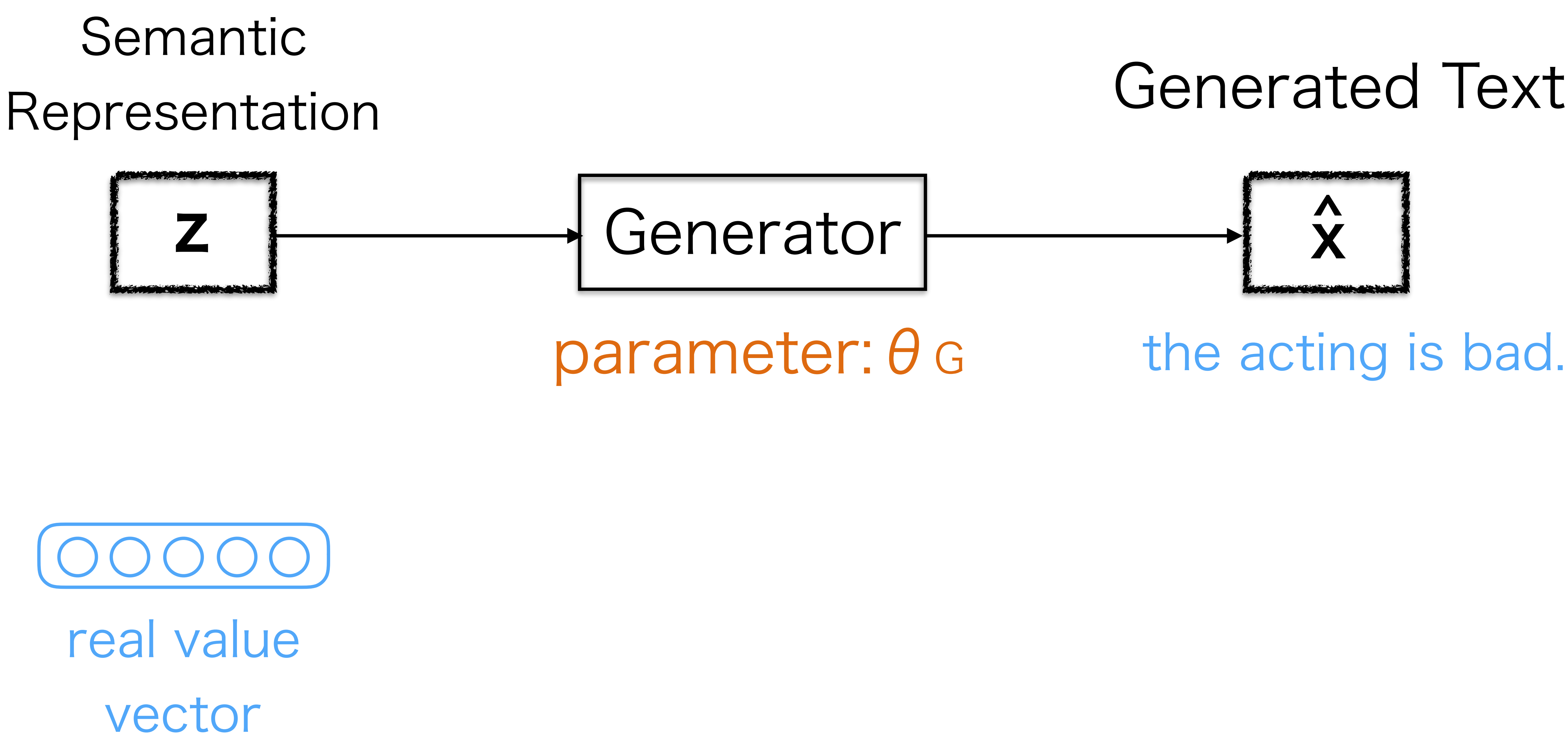
厳密性を欠く、ゆるふわな説明

イメージだけでもつかんでもらえれば

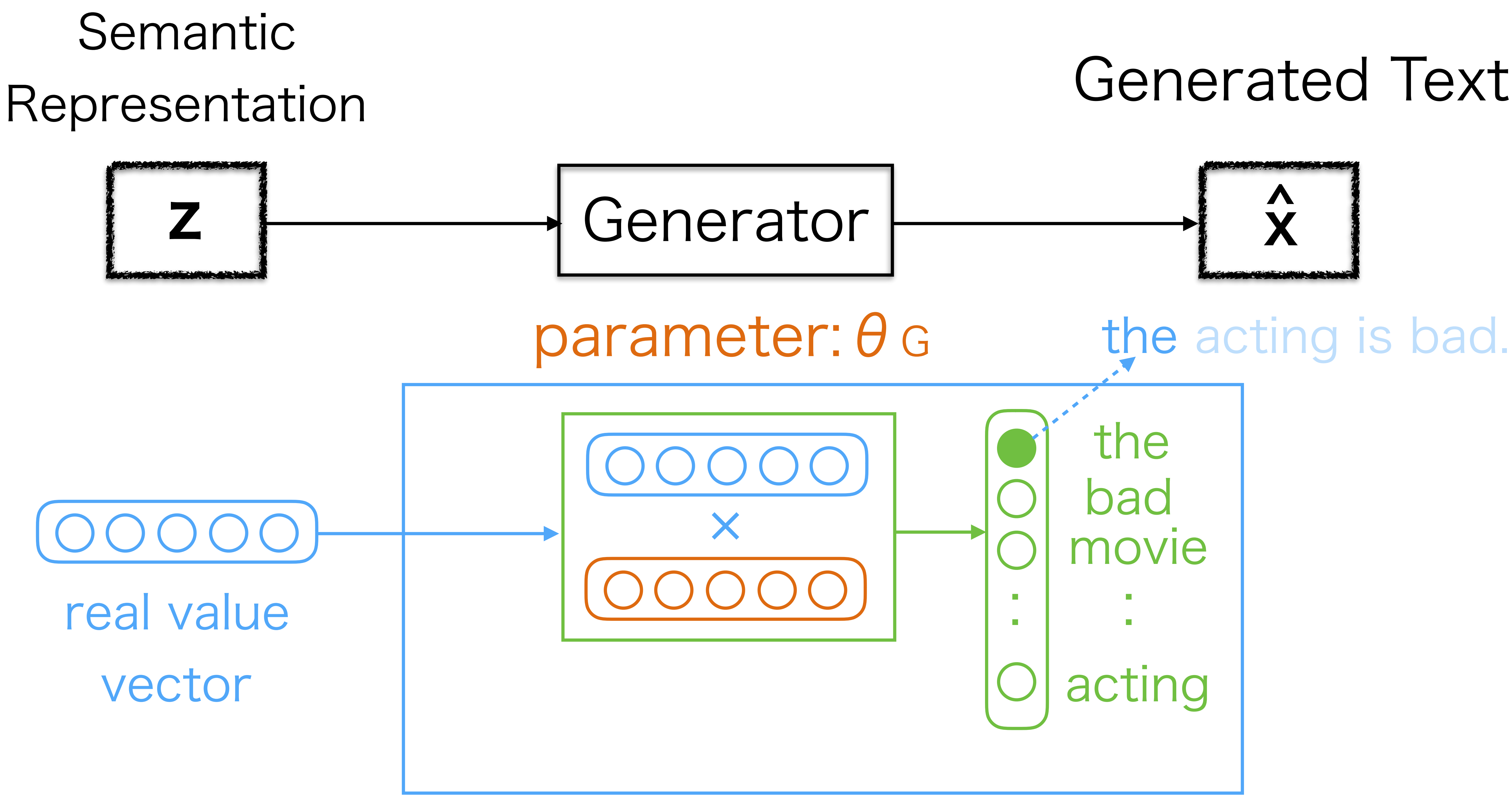
How to Generate Text?



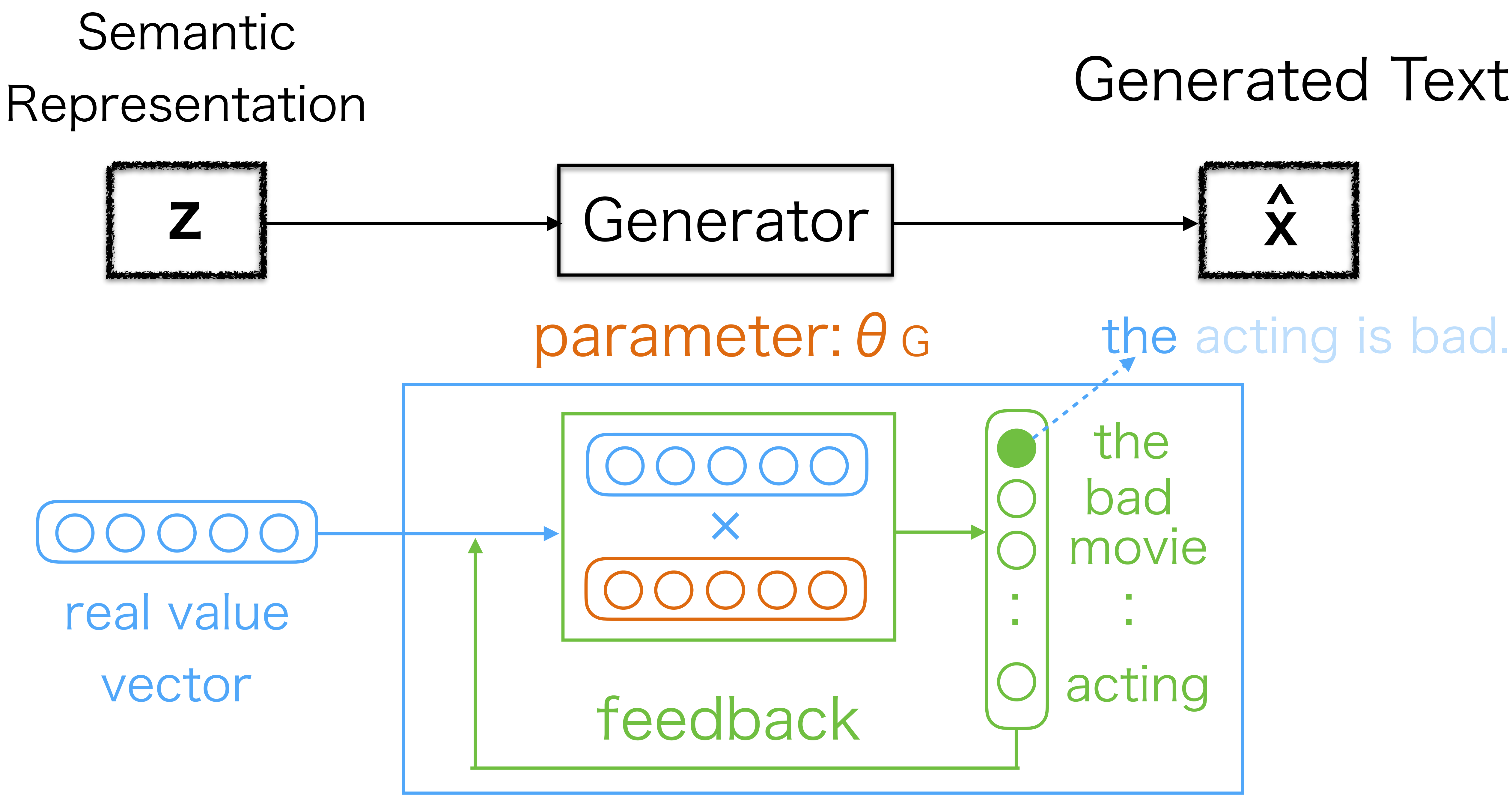
How to Generate Text?



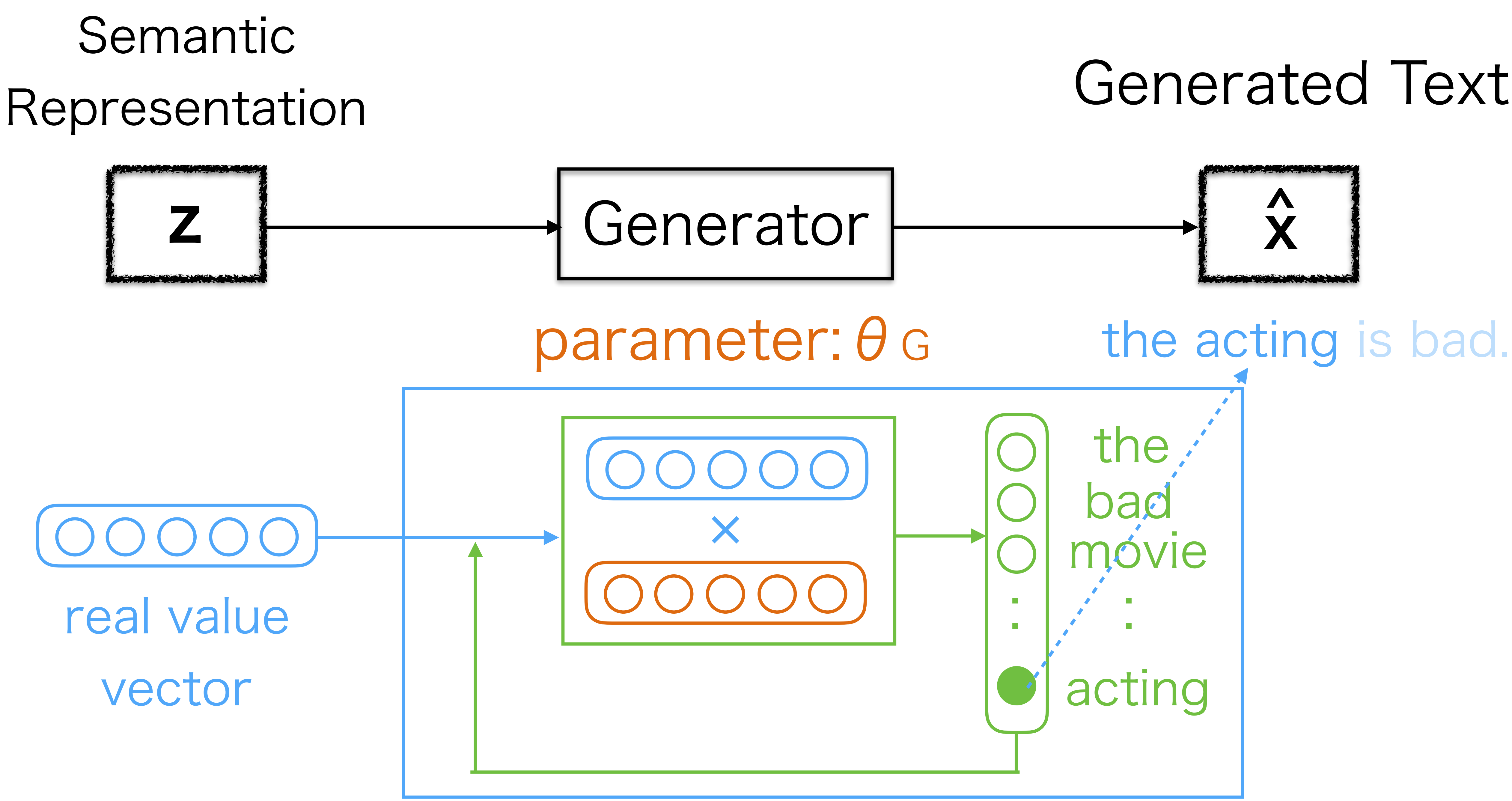
How to Generate Text?



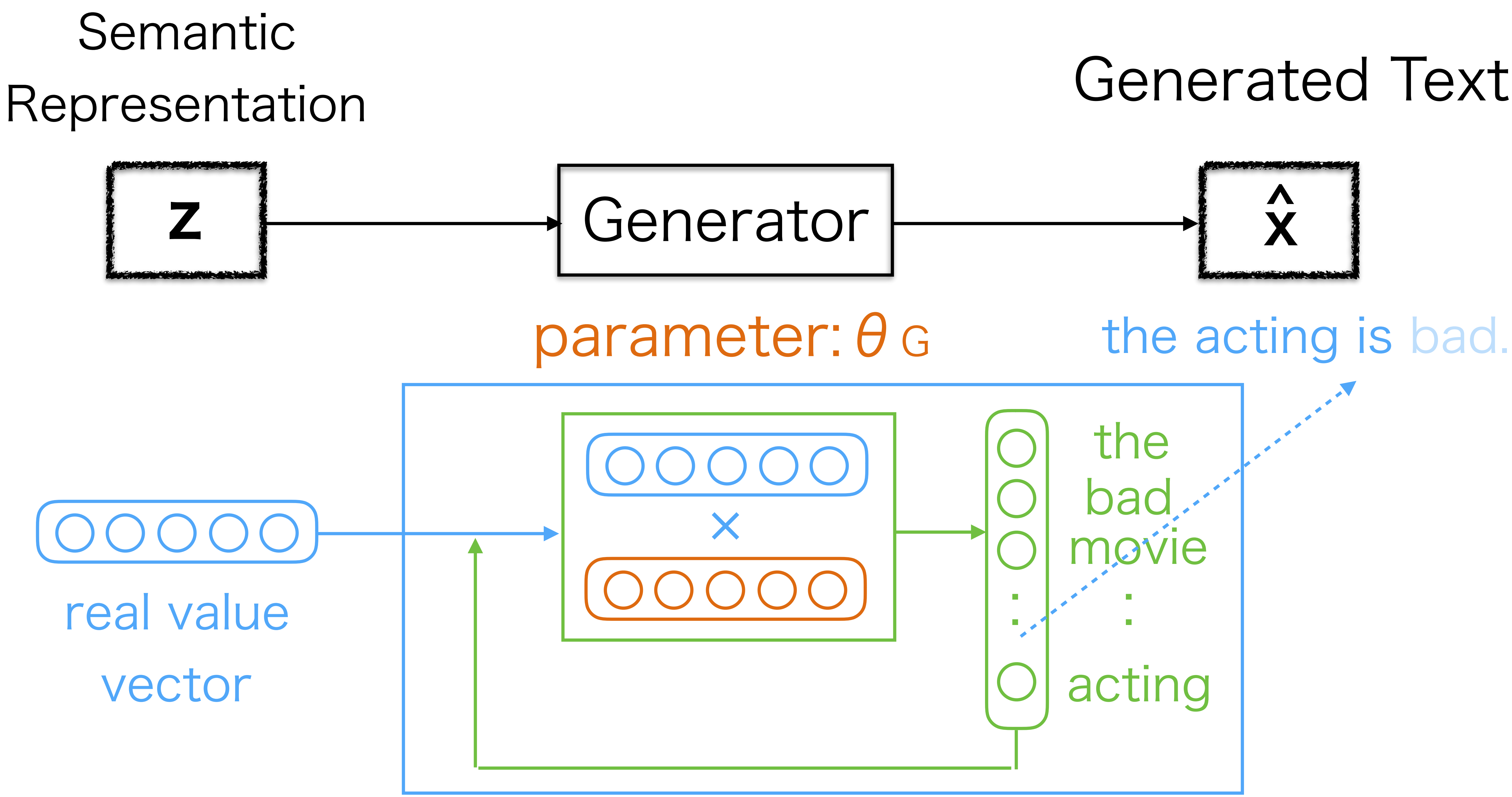
How to Generate Text?



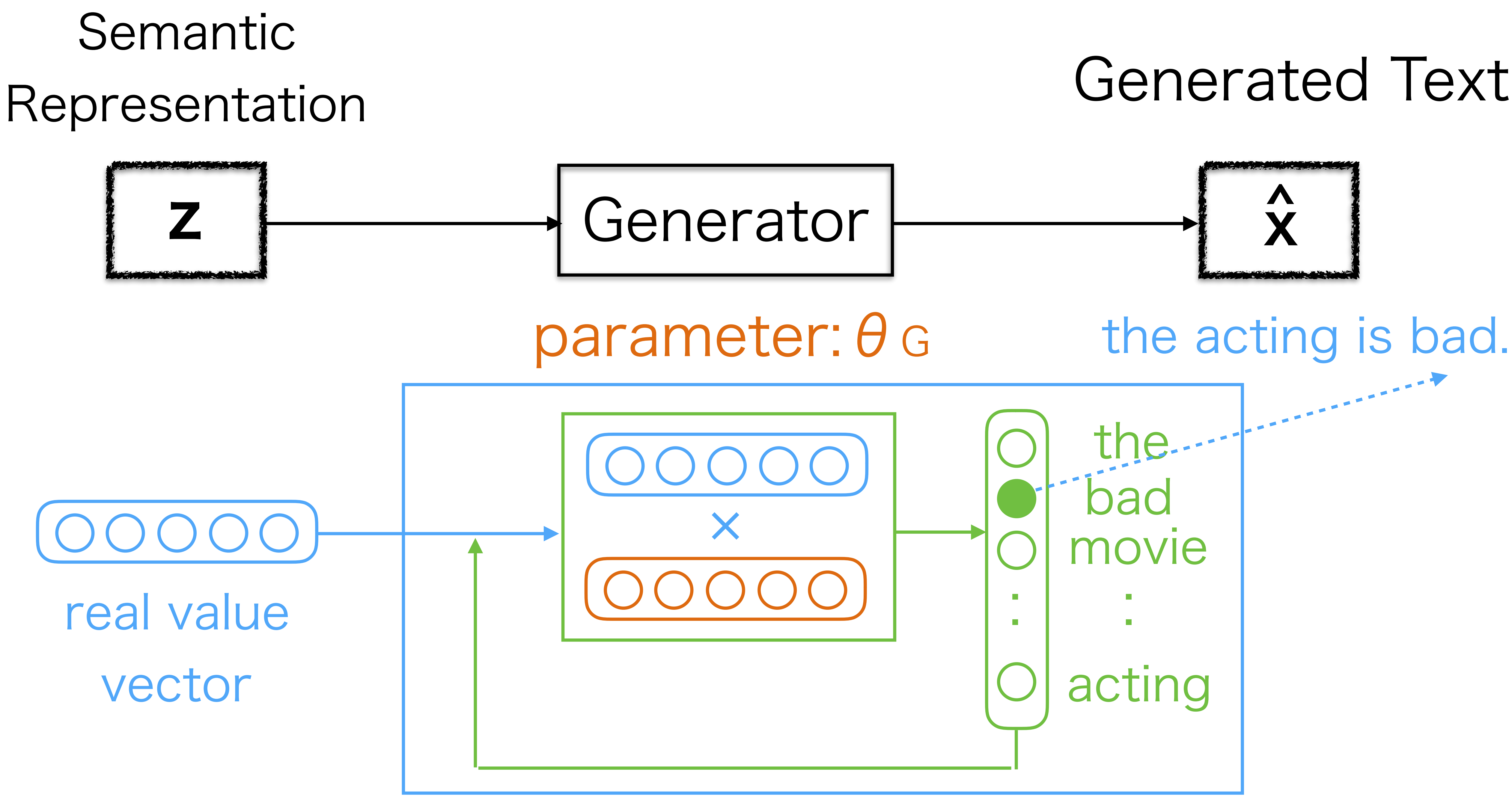
How to Generate Text?



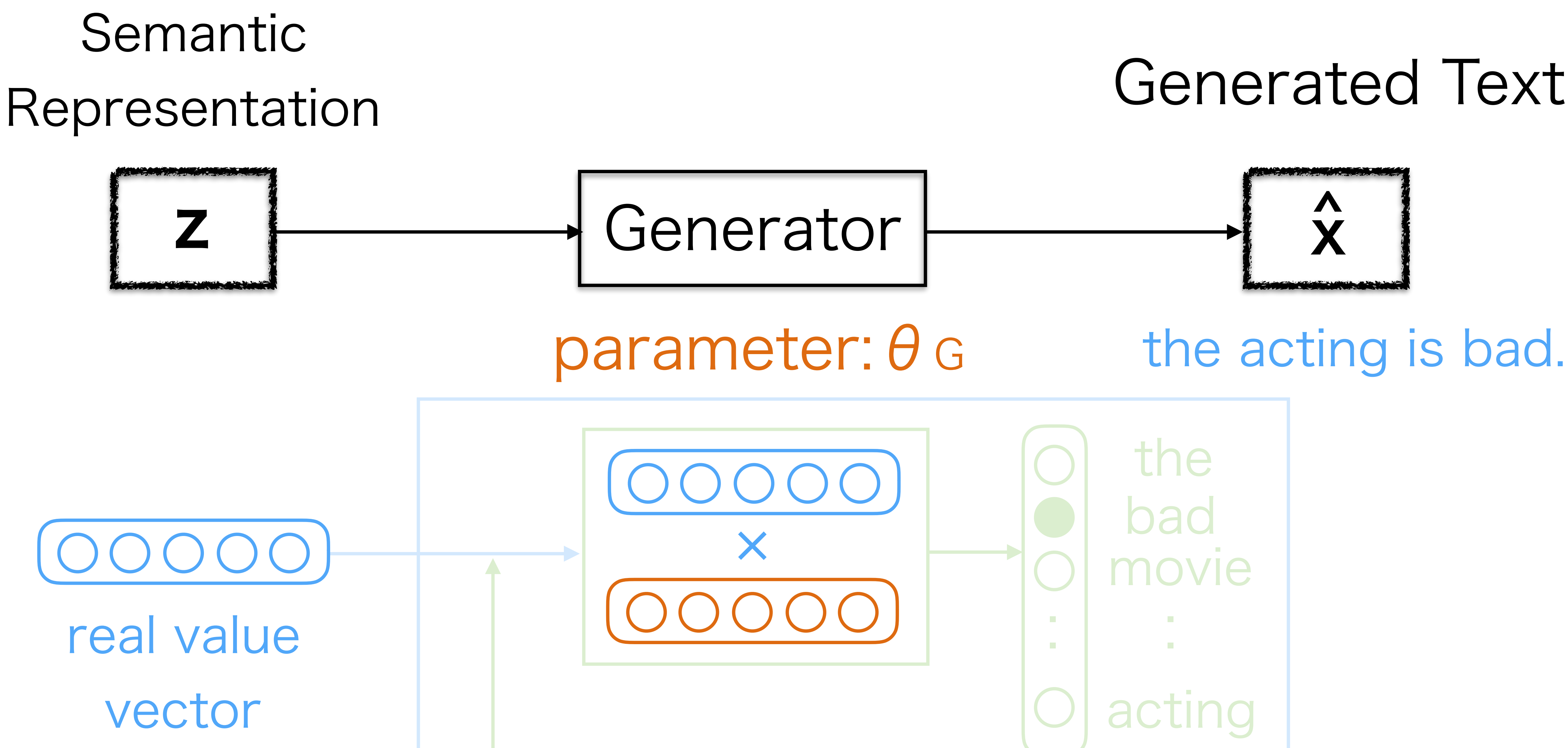
How to Generate Text?



How to Generate Text?

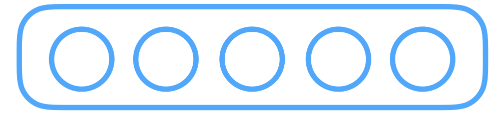


How to Generate Text?

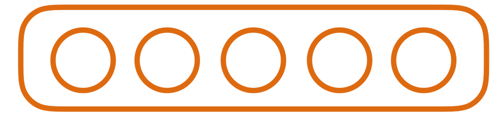


1. We want to get adequate and .
2. **Learning 1 from real text data.**

semantic represetation

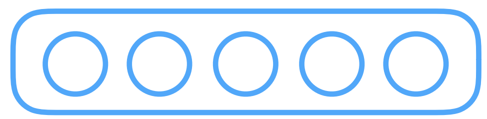


parameter θ_G

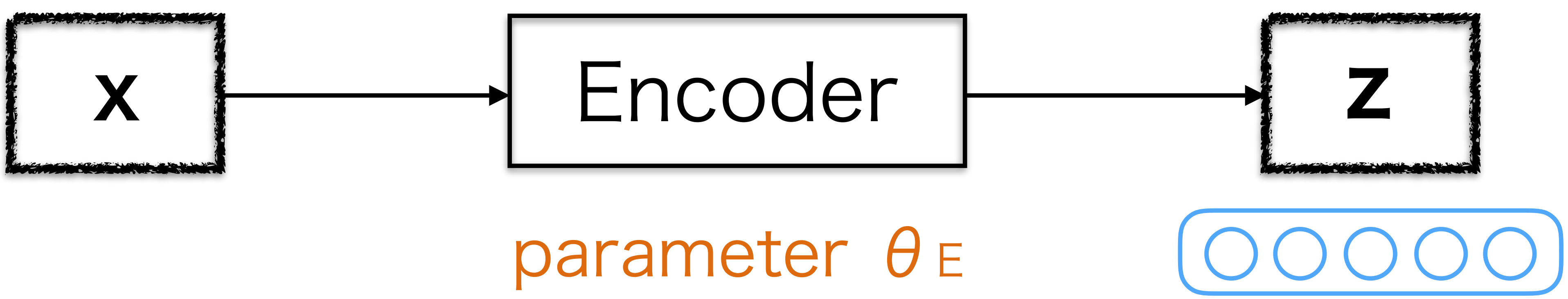


How to get ... from text data ?

We have text data **{x}**

Idea 1: Estimate  from **{x}**.

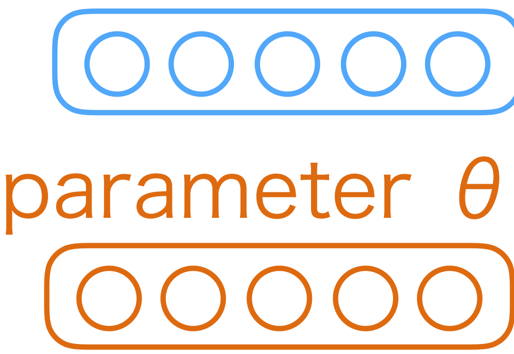
the acting is bad



How to get

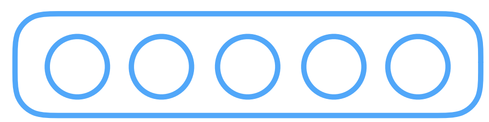
semantic represetation

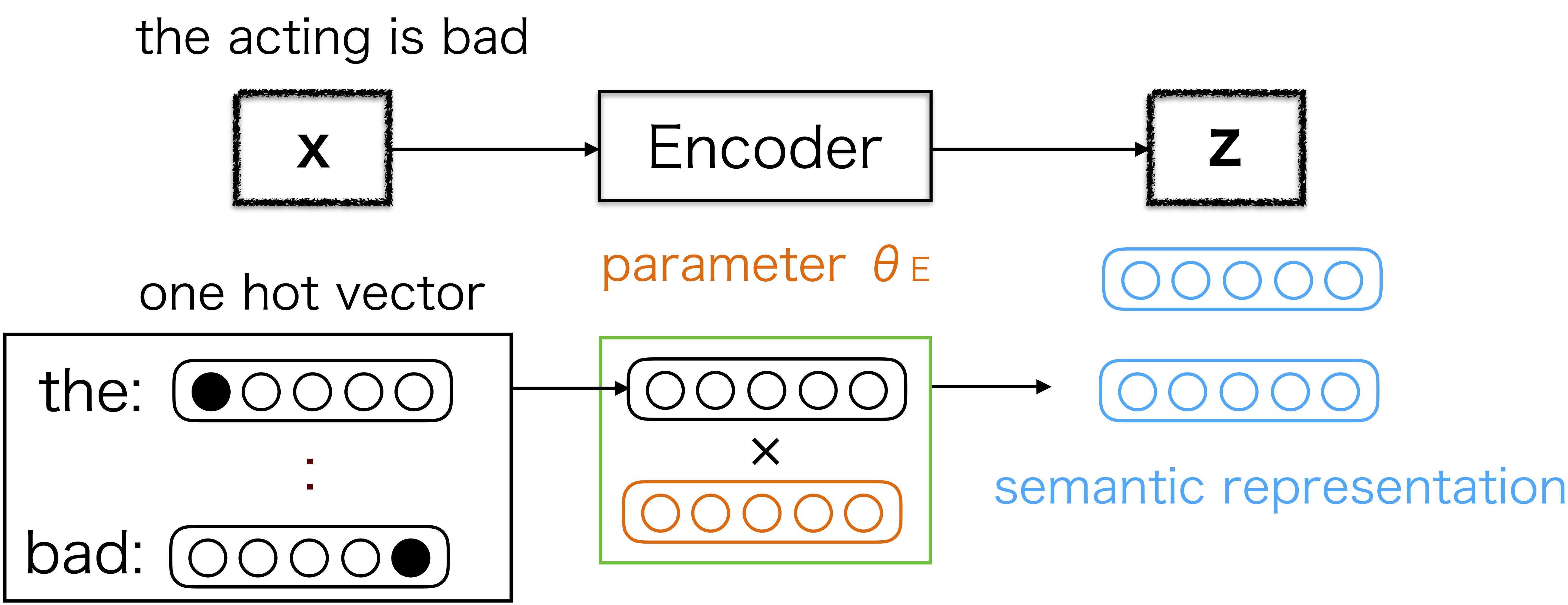
parameter θ_G

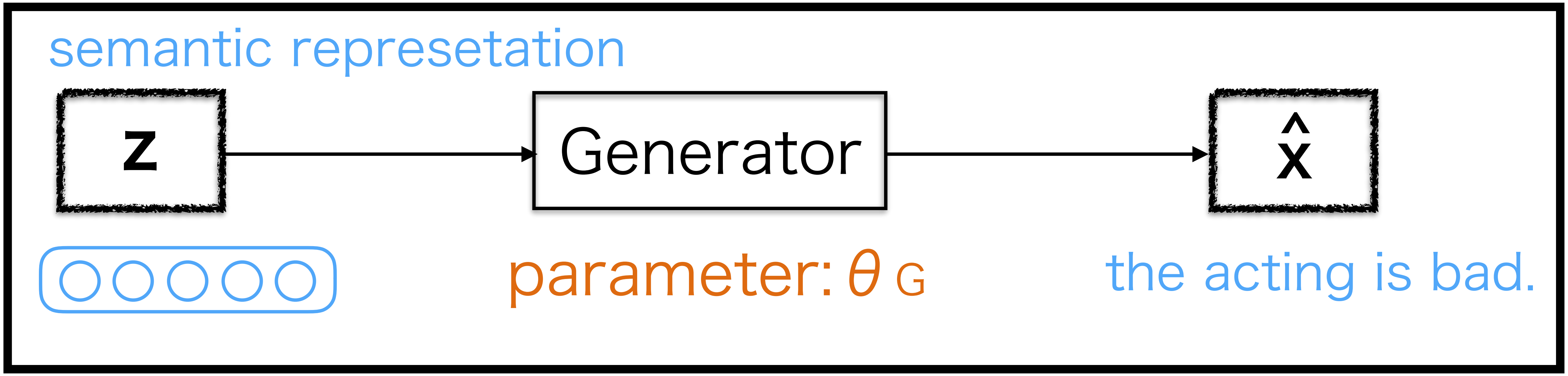


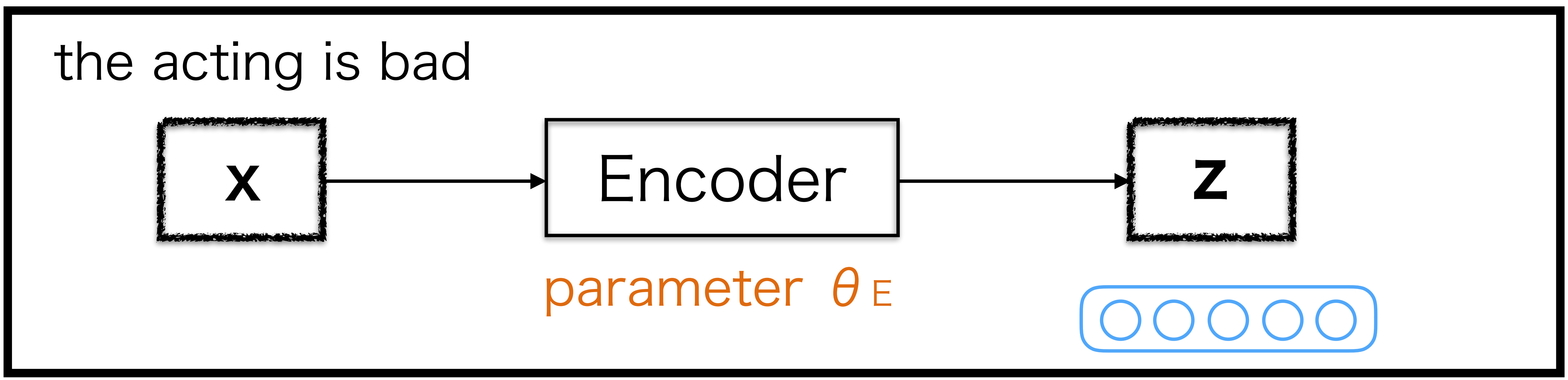
from text data ?

We have text data **{x}**

Idea: Estimate  from **{x}**.

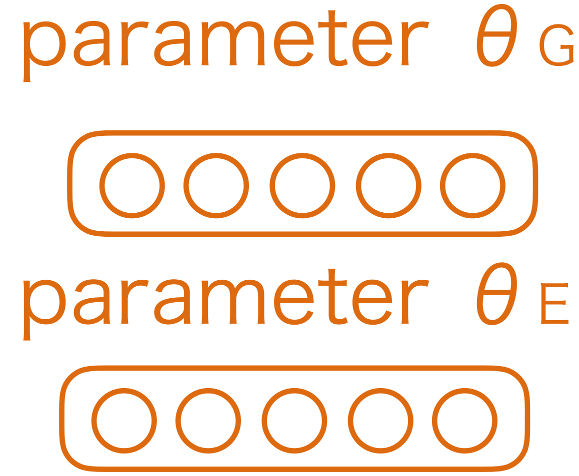






Goal: Automatically learn

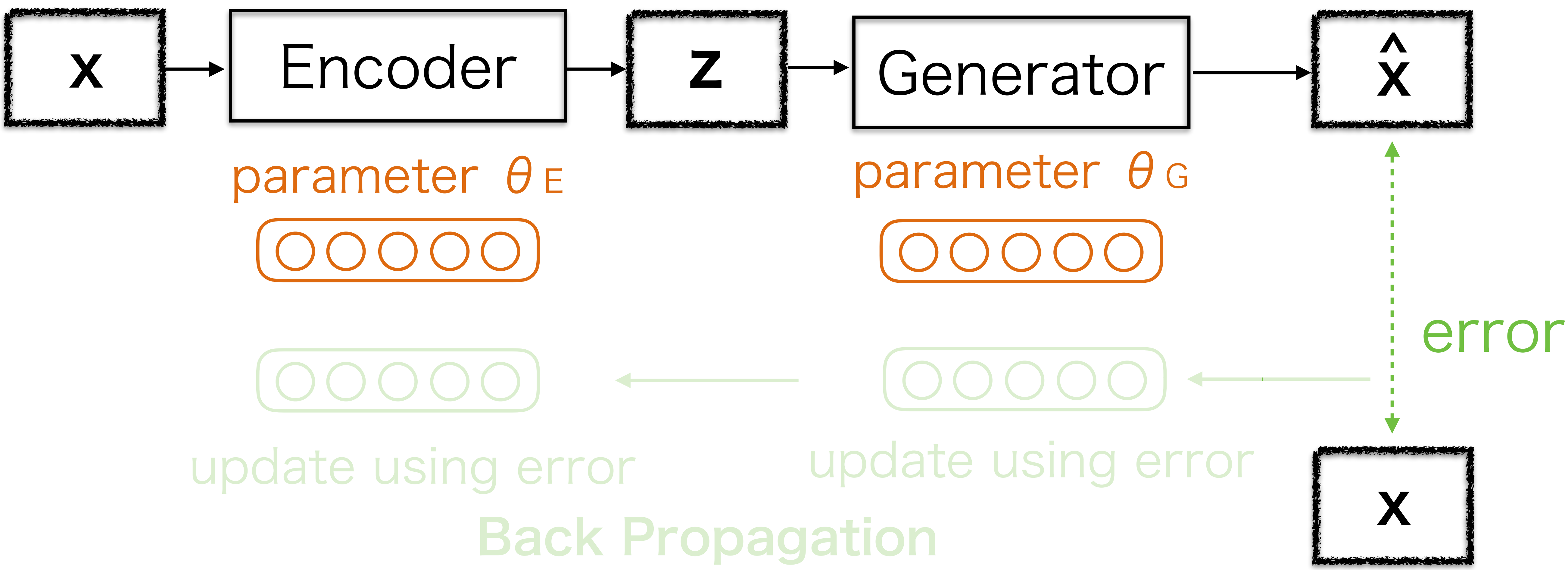
How to get



?

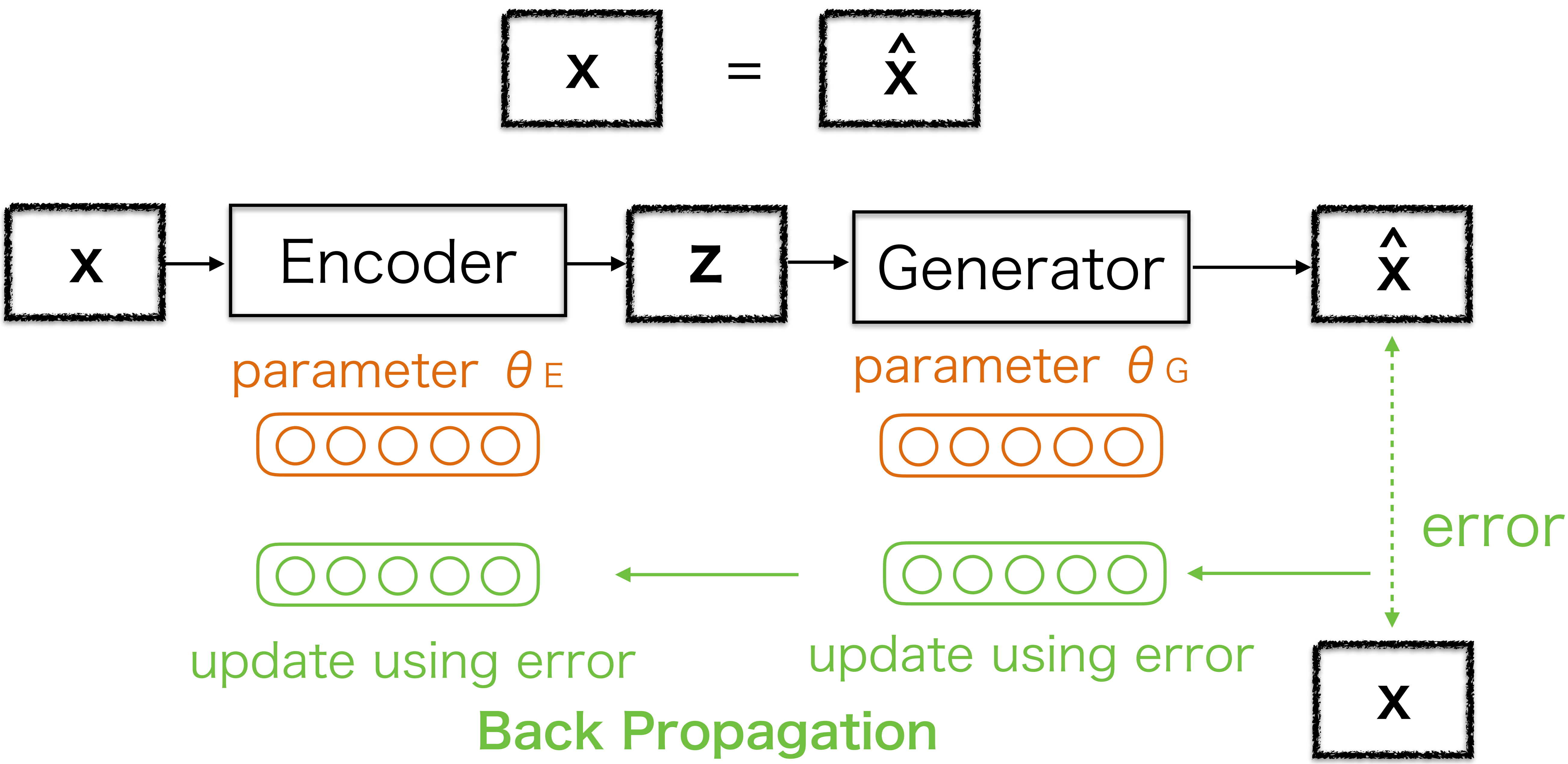
Idea 2: Learn to minimize reconstruction error

$$x = \hat{x}$$



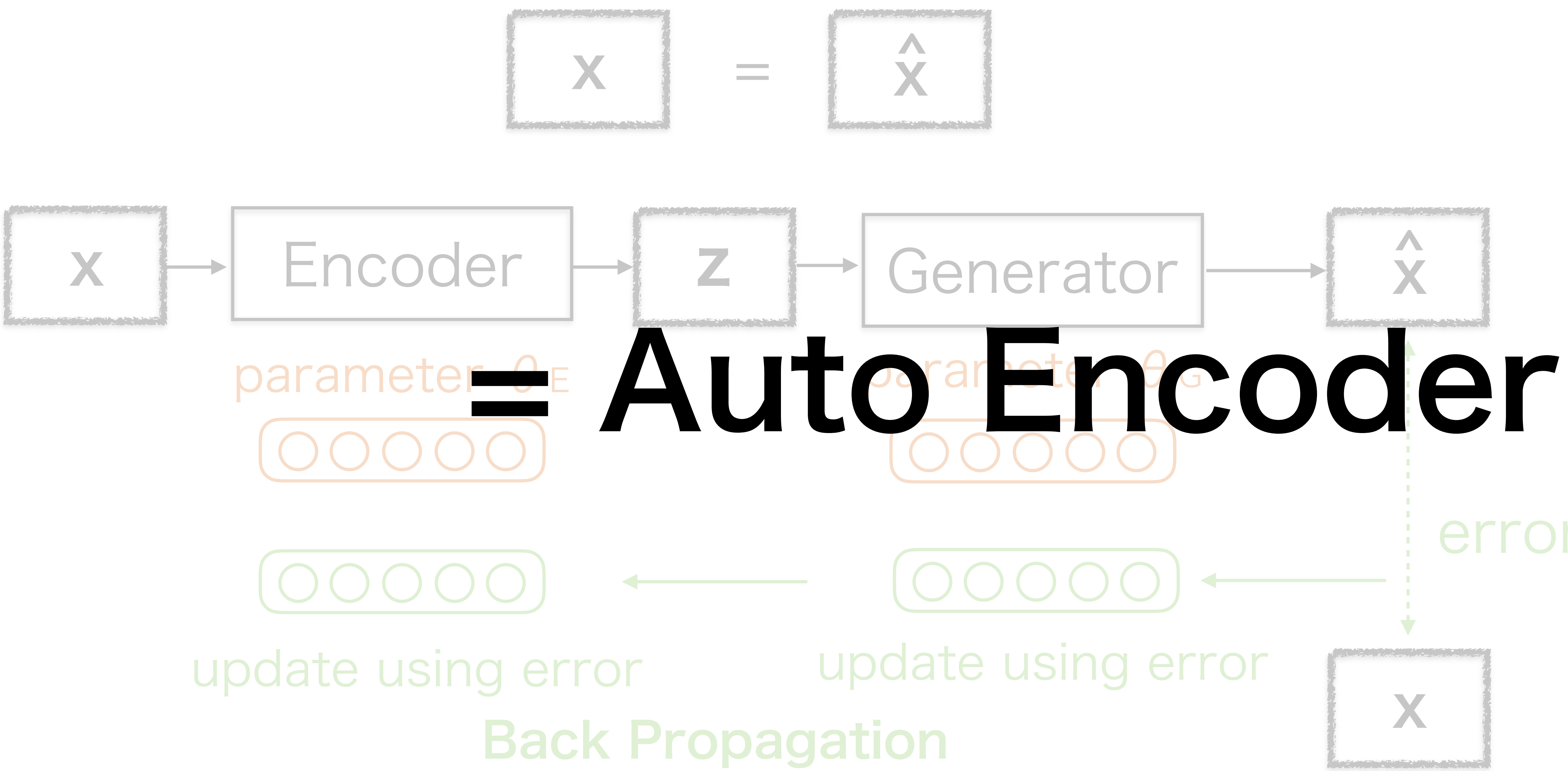
How to get parameter θ_G parameter θ_E ?

Idea 2: Learn to minimize reconstruction error

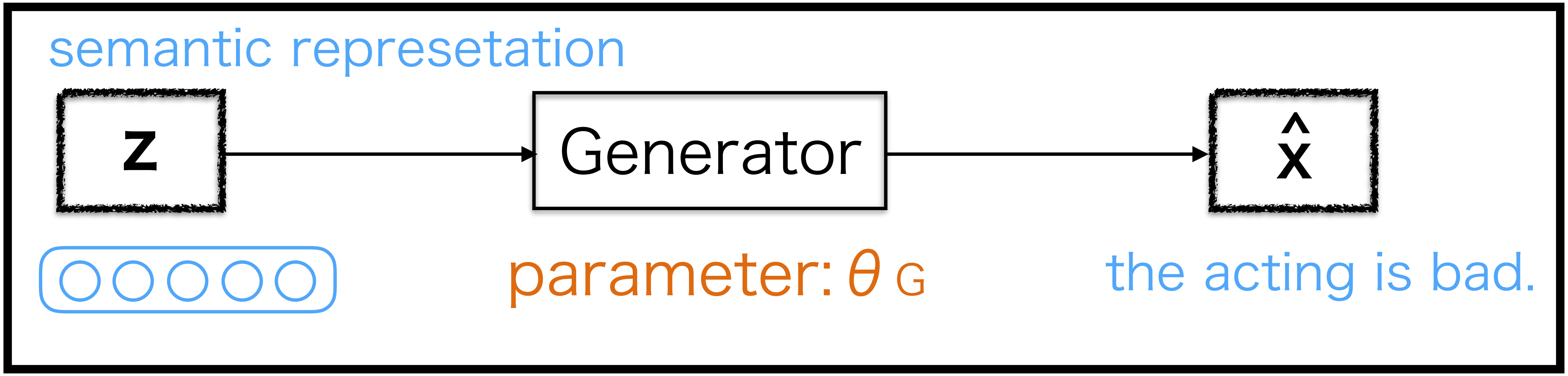


How to get parameter θ_G parameter θ_E ?

Idea 2: Learn to minimize reconstruction error



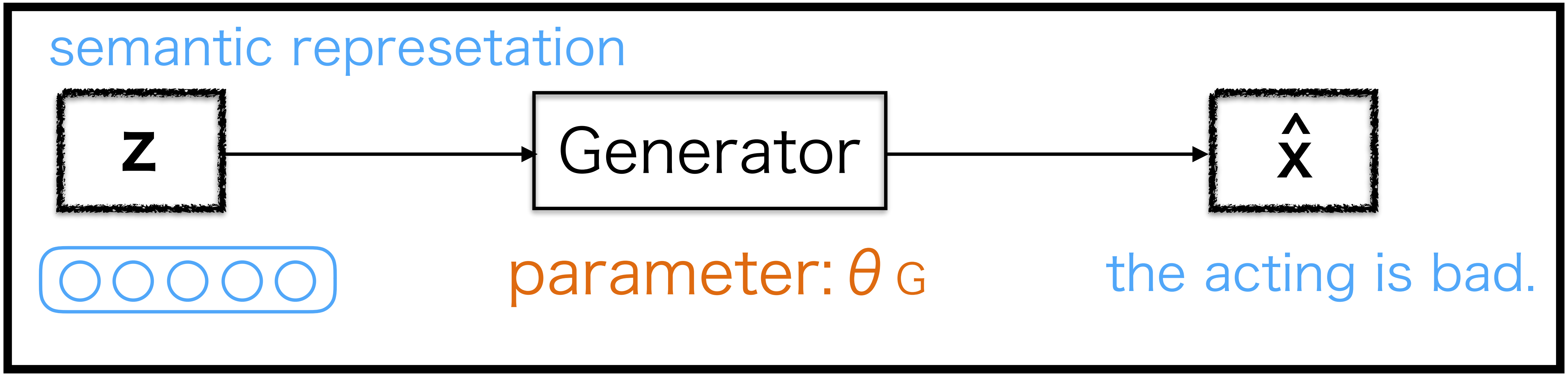
= Auto Encoder



If we set **z** (from prior distribution p(**z**))

we can generate $\hat{\mathbf{x}}$ according to **z**

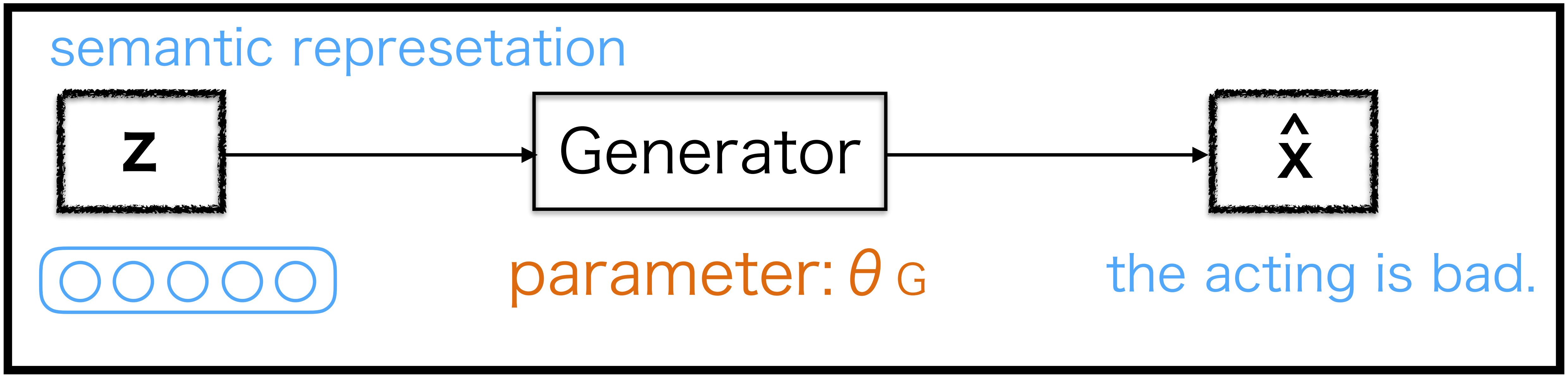
Problem



1. How can we generate texts we want to.

   We don't know how to set **z** to generate texts.

z is hard to interpret by human

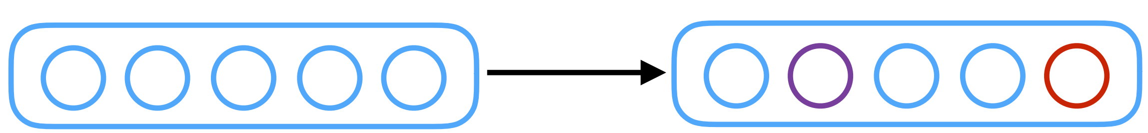


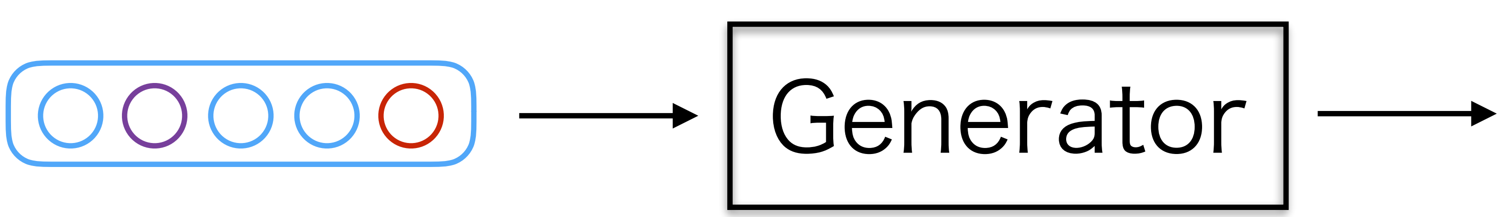
Given:

If we want to generate text “the acting is good”

We know semantic representation ○○○○○ of “the acting is bad”

How to generate it?

e.x. manipulate 



this is one of the best films

Characteristic of this work:

1. Enable to control attribute of Generated Text

the acting is bad the acting is good	e.g. attribute -> sentiment

i thought the movie was good i guess the movie is good i guess the movie will have been good	e.g. attribute -> tense

2. Propose learning technique to controllable generation

   Variational Auto Encoder (VAE) + attribute discriminator

Task: Text Generation

Characteristic of this work:

1. Enable to control attribute of Generated Text

the acting is bad the acting is good	e.g. attribute -> sentiment

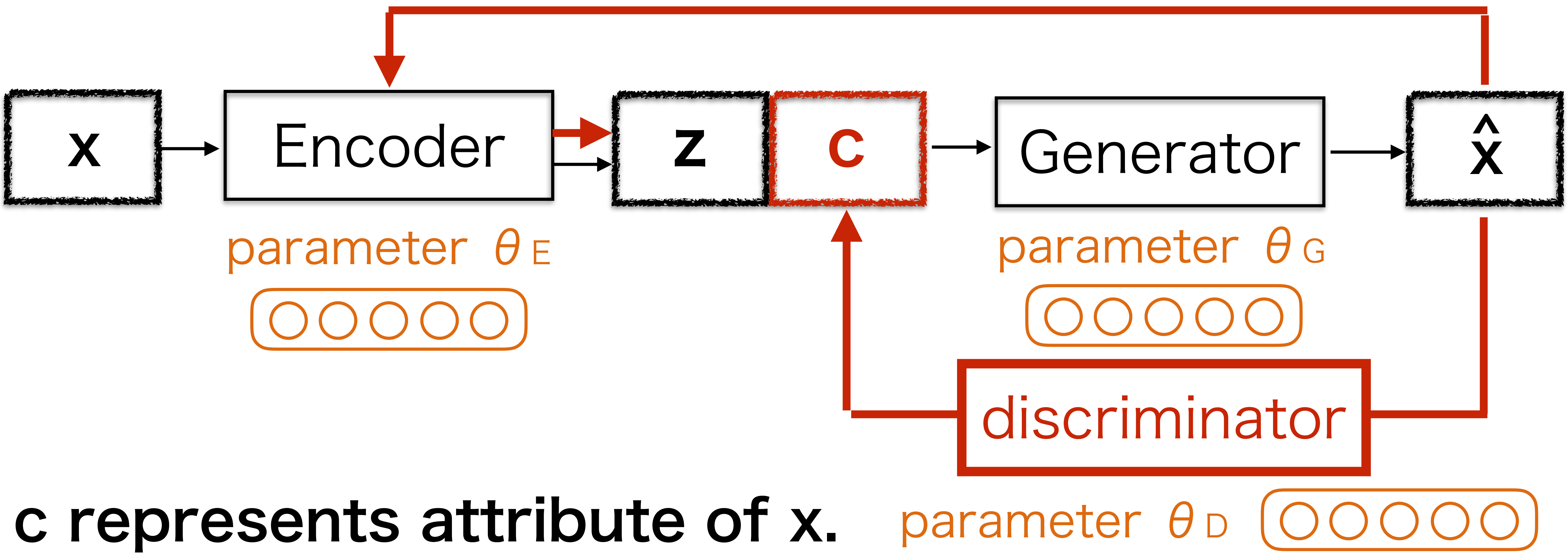
i thought the movie was good i guess the movie is good i guess the movie will have been good	e.g. attribute -> tense

2. Propose learning technique to controllable generation

Variational Auto Encoder (VAE) + attribute discriminator

Task: Text Generation

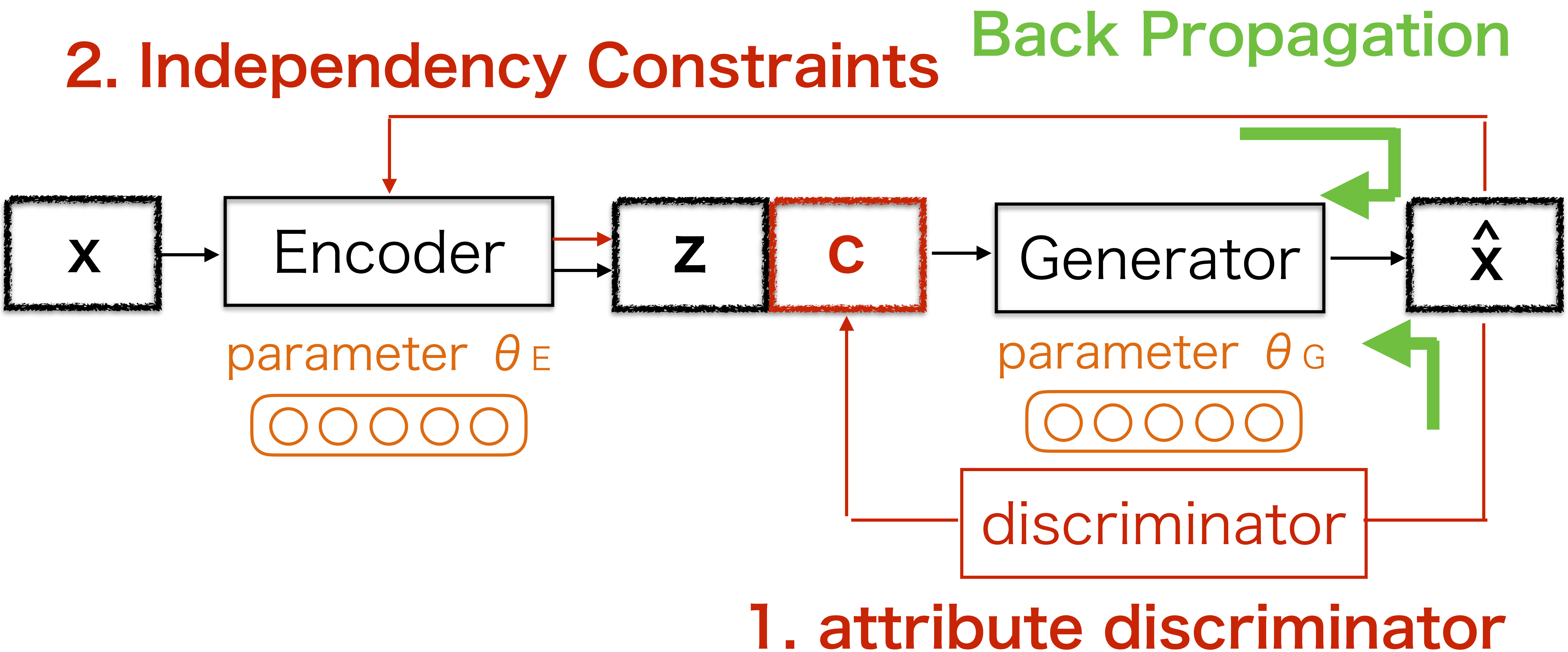
Overview of the framework



c represents attribute of x.

1. Add discriminator that classify attribute c of $\hat{x}$
2. Regard Encoder as another discriminator that classify **z** of $\hat{x}$

Overview of the framework



1. Learning to generate $\hat{\mathbf{x}}$ from specific attribute **c**
2. Ensure independency between **z** and **c**

Encoder and Generator

Encoder:

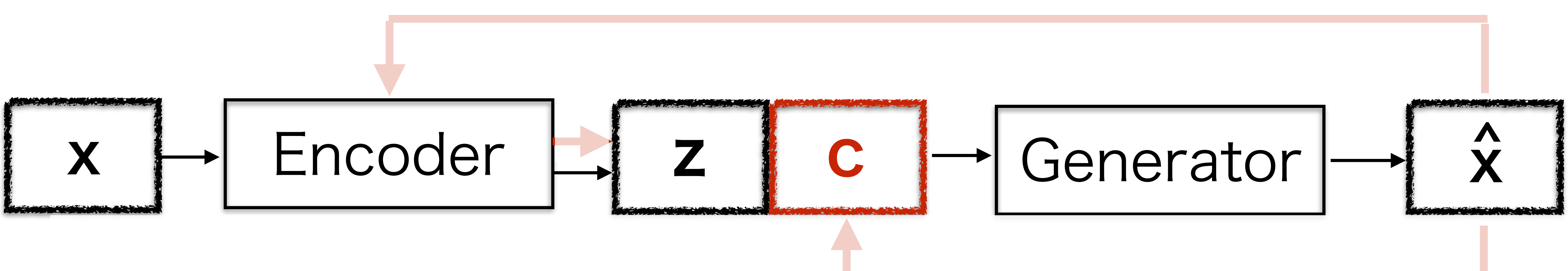
$$z \sim E(x) = q_E(z|x).$$

Generator:

$$\hat{x} \sim G(z, c) = p_G(\hat{x}|z, c) = \prod_t p(\hat{x}_t|\hat{x}^{<t}, z, c),$$

$$\hat{x}_t \sim \text{softmax}(o_t/\tau),$$

- Using LSTM as Encoder and Decoder
- **z**: from Gaussian prior p(**z**)
  c: from categorical distribution p(**c**)



Learning Encoder and Generator

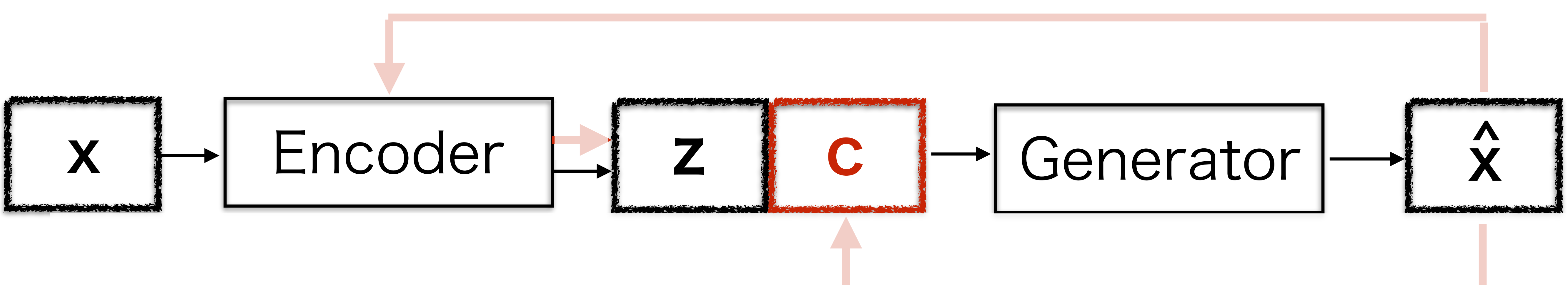
Loss Function 1:

$$\mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_E; \boldsymbol{x}) = -\mathrm{KL}(q_E(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})) + \mathbb{E}_{q_E(\boldsymbol{z}|\boldsymbol{x}) q_D(\mathbf{c}|\boldsymbol{x})} \left[\log p_G(\boldsymbol{x}|\boldsymbol{z}, \mathbf{c})\right],$$

regularization

lower bound of log likelihood

- Maximize the lower bound of log-likelihood



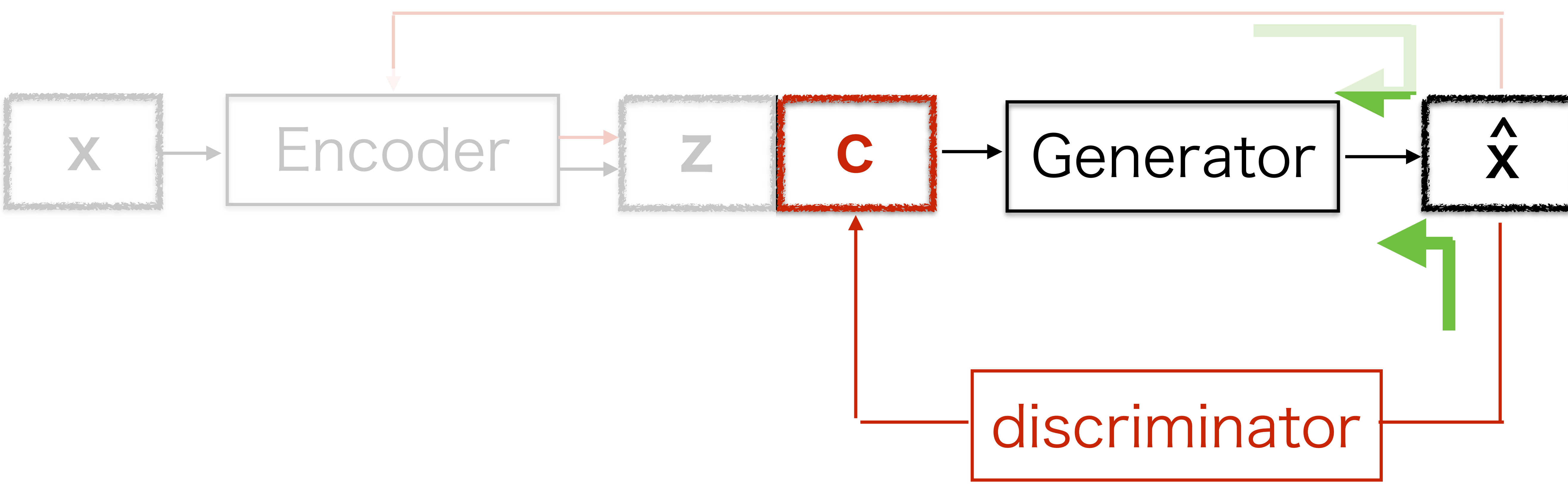
Learning Generator

Discriminator:

$$D(\boldsymbol{x}) = q_D(\boldsymbol{c}|\boldsymbol{x}).$$ (Using Convolutional Neural Network (CNN))

Loss Function 2:

$$\mathcal{L}_{\text{Attr},c}(\boldsymbol{\theta}_G) = \mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})}\left[\log q_D(\boldsymbol{c}|\widetilde{G}_\tau(\boldsymbol{z},\boldsymbol{c}))\right].$$

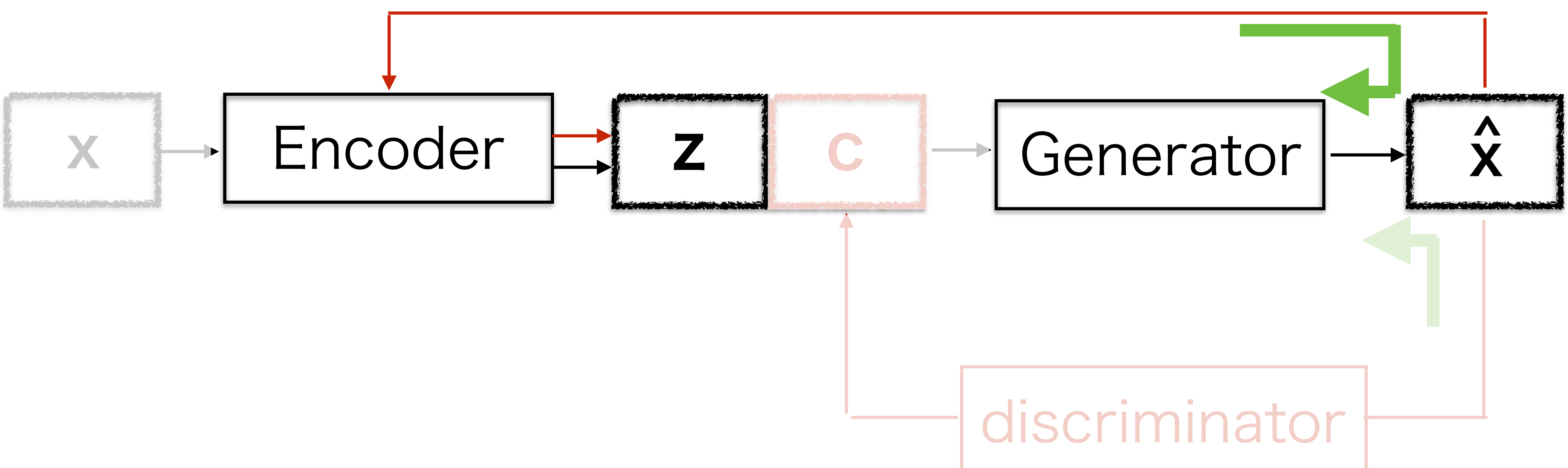


Learning Generator

Loss Function 3:

- Independency Constraints

$$\mathcal{L}_{\text{Attr},z}(\boldsymbol{\theta}_G) = \mathbb{E}_{p(\boldsymbol{z})p(\boldsymbol{c})}\left[\log q_E(\boldsymbol{z}|\widetilde{G}_\tau(\boldsymbol{z},\boldsymbol{c}))\right].$$

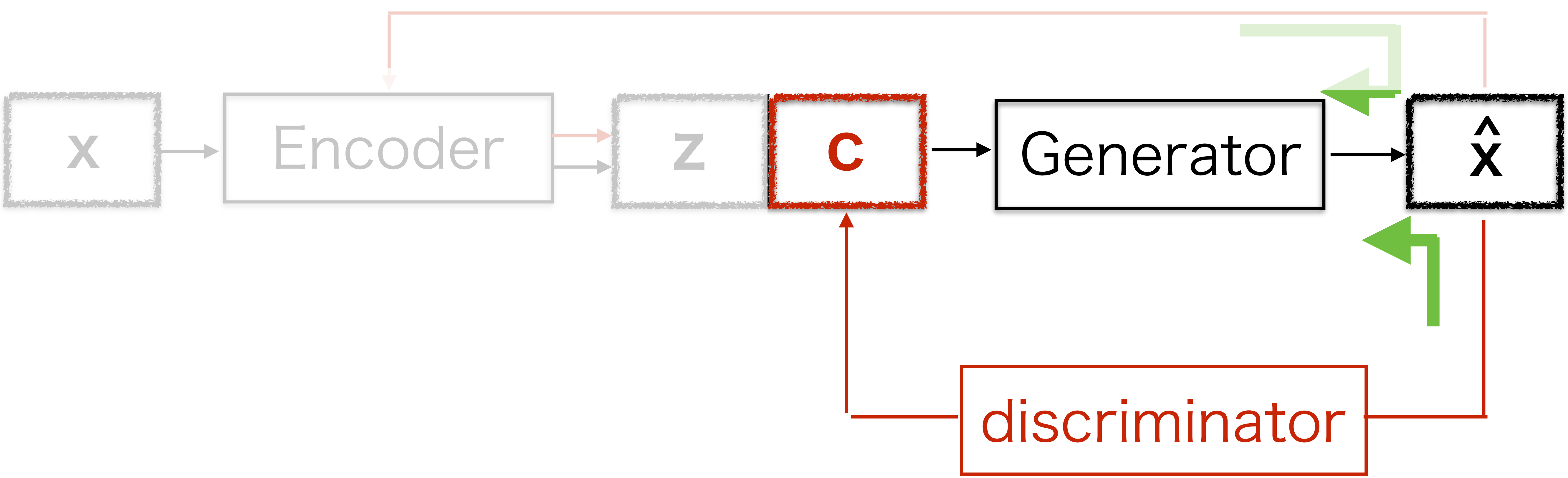


Learning Discriminator

- Using Labeled data: $\mathcal{X}_L = \{(\boldsymbol{x}_L, \boldsymbol{c}_L)\}$
- Learning to predict true label from XL

Loss Function 1:

$$\mathcal{L}_s(\boldsymbol{\theta}_D) = \mathbb{E}_{\mathcal{X}_L}\left[\log q_D(\boldsymbol{c}_L|\boldsymbol{x}_L)\right].$$

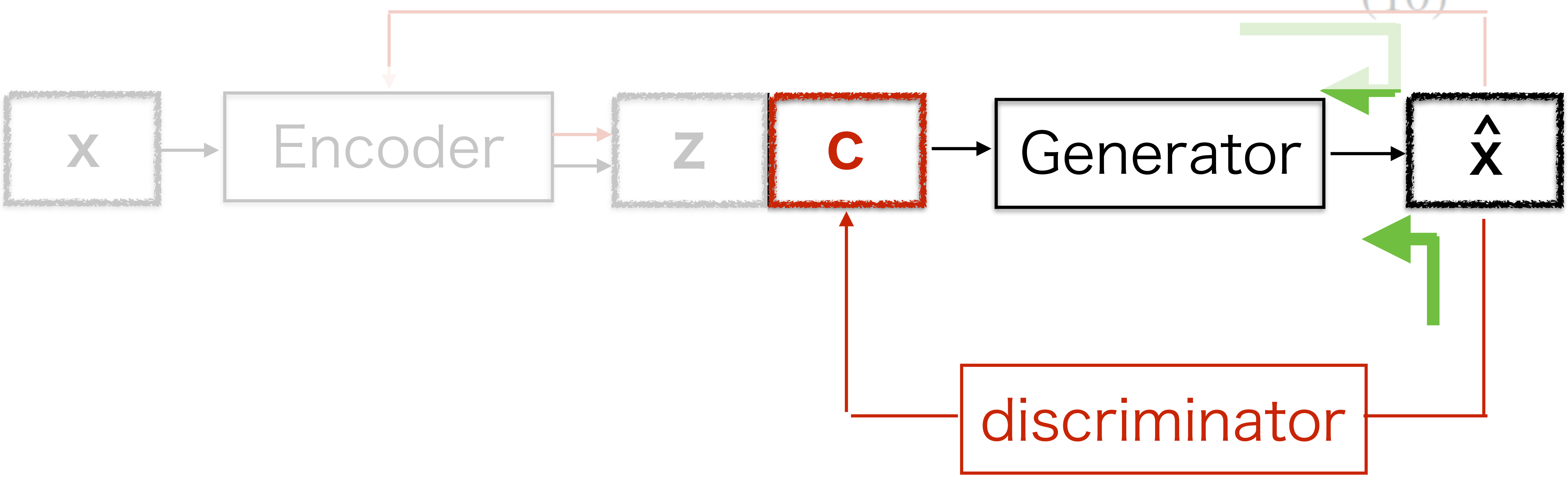


Learning Discriminator

- Using generated text $\hat{\mathbf{x}}$ from **c**

Loss Function 2:

$$\mathcal{L}_u(\boldsymbol{\theta}_D) = \mathbb{E}_{p_G(\hat{\boldsymbol{x}}|\boldsymbol{z},\boldsymbol{c})p(\boldsymbol{z})p(\boldsymbol{c})}\left[\log q_D(\boldsymbol{c}|\hat{\boldsymbol{x}}) + \beta\mathcal{H}(q_D(\boldsymbol{c}'|\hat{\boldsymbol{x}}))\right], \quad (10)$$



Learning Procedure

1. Learning VAE using Large Unlabeled Data

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_E; \boldsymbol{x}) = -\,\text{KL}(q_E(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})) + \mathbb{E}_{q_E(\boldsymbol{z}|\boldsymbol{x}) q_D(\boldsymbol{c}|\boldsymbol{x})} \left[\log p_G(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c}) \right]$$

2. Learning Discriminator and VAE alternately

I. Learning Discriminator

$$\min_{\boldsymbol{\theta}_D} \mathcal{L}_D = \mathcal{L}_s + \lambda_u \mathcal{L}_u,$$

II. Learning Encoder and Generator

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}_G, \boldsymbol{\theta}_E; \boldsymbol{x}) : \quad \min_{\boldsymbol{\theta}_G} \mathcal{L}_G = \mathcal{L}_{\text{VAE}} + \lambda_c \mathcal{L}_{\text{Attr},c} + \lambda_z \mathcal{L}_{\text{Attr},z},$$

Experiments: Dataset

- **Unlabeled data (to train autoencoder):**
  - IMDB data: Reviews of movies
    1.4M sentences, vocabulary size 16K
- **Labeled data (with attribute):**
  - sentiment label: {positive, negative}
    - Stanford Sentiment Treebank
      - SST full: 2,837/872/1821 (train/dev/test)
      - SST small: 250/872/1821 (train/dev/test)
    - Lexicon: 2700 words with sentiment labels [Wilson et al. 2005]
    - IMDB 5K/1K/10K sentences
  - (tense: {past, present, future})
    (- 5250 words with tense labels from timebank)

Experimental Results: sentiment

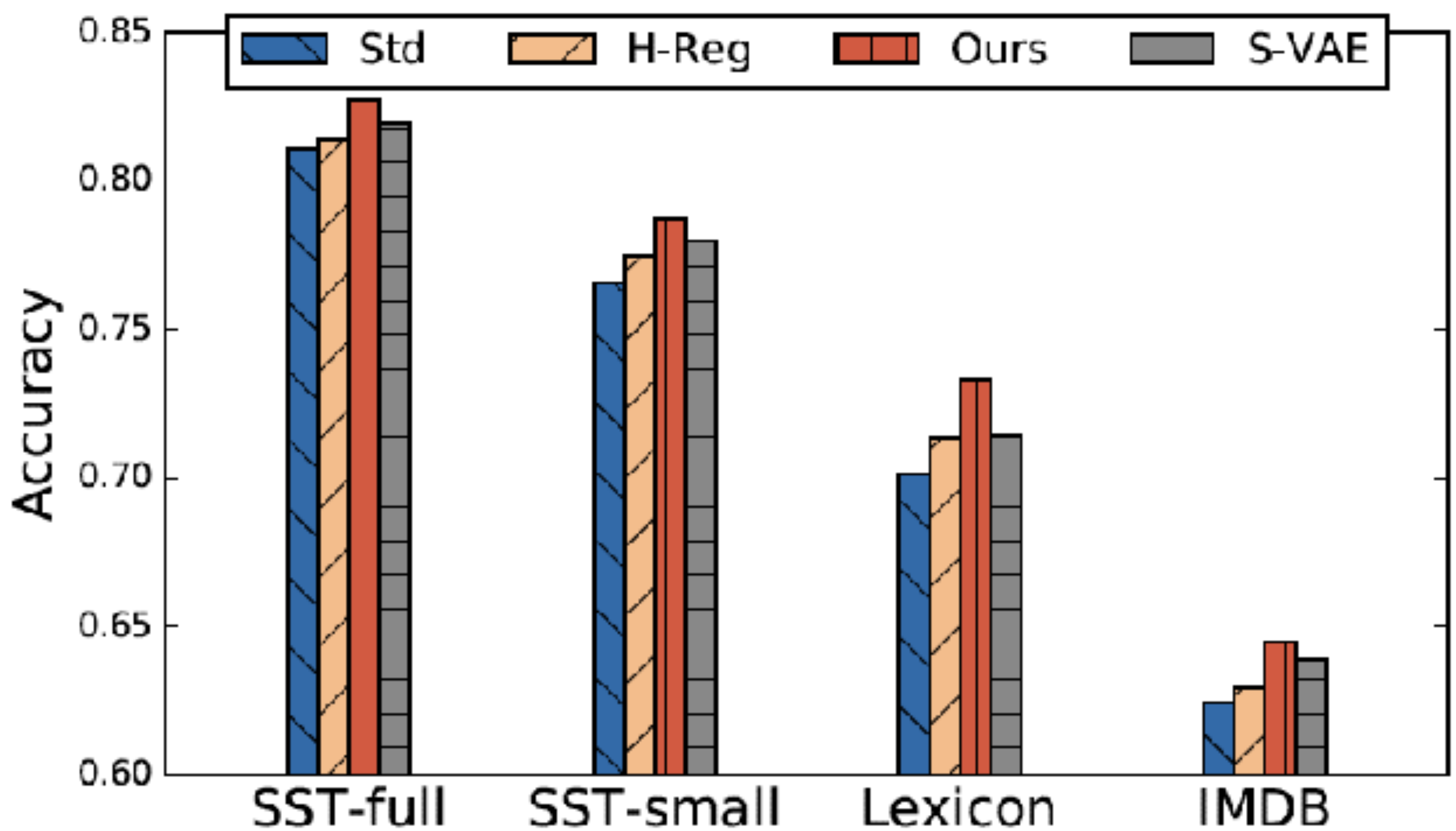
- Generate sentences from true **c** and predict label from generated sentences using s.o.t.a sentiment classifier
- Metric: Accuracy

Model	Dataset		
	SST-full	SST-small	Lexicon
S-VAE	0.822	0.679	0.660
Ours	**0.851**	**0.707**	**0.701**

Table 1. Accuracy of generated sentences measured by a pre-trained sentiment classifier (Hu et al., 2016a). Models are trained on the three sentiment datasets and generate 30K sentences, respectively. S-VAE denotes the semi-supervised VAE model (Kingma et al., 2014).

Experimental Results: Augument Dataset

- Generate sentences from proposed method and S-VAE
- Add generated sentences to training data
- Training sentiment classifier using augmented training data



Experimental Result:

fix unstructure code **z**

only change attribute code **c**

w/ independency constraint	w/o independency constraint
the film is strictly routine !	the acting is bad .
the film is full of imagination .	the movie is so much fun .
after watching this movie , i felt that disappointed .	none of this is very original .
after seeing this film , i 'm a fan .	highly recommended viewing for its courage , and ideas .
the acting is uniformly bad either .	too bland
the performances are uniformly good .	highly watchable
this is just awful .	i can analyze this movie without more than three words .
this is pure genius .	i highly recommend this film to anyone who appreciates music .
nothing about this film is amazing	a movie version of a paint by numbers
nothing about this film is terrible	a backstage must see for true fans of the bard

Experimental Result:

fix unstructure code **z**

only change attribute code **c:**

attribute -> sentiment and tense

Varying the code of sentiment	Varying the code of tense
this movie was awful and boring .	this was one of the outstanding thrillers of the last decade
this movie was funny and touching .	this is one of the outstanding thrillers of the all time
	this will be one of the great thrillers of the all time
jackson is n't very good with documentary	
jackson is superb as a documentary productions	i thought the movie was too bland and too much
	i guess the movie is too bland and too much
you will regret it	i guess the film will have been too bland
you will enjoy it	

Experimental Result:

fix attribute code **c**
only change unstructured code **z**

Varying the unstructured code z	
("negative", "past")	*("positive", "past")*
the acting was also kind of hit or miss .	his acting was impeccable
i wish i 'd never seen it	this was spectacular , i saw it in theaters twice
by the end i was so lost i just did n't care anymore	it was a lot of fun
("negative", "present")	*("positive", "present")*
the movie is very close to the show in plot and characters	this is one of the better dance films
the era seems impossibly distant	i 've always been a big fan of the smart dialogue .
i think by the end of the film , it has confused itself	i recommend you go see this, especially if you hurt
("negative", "future")	*("positive", "future")*
i wo n't watch the movie	i hope he 'll make more movies in the future
and that would be devastating !	i will definitely be buying this on dvd
i wo n't get into the story because there really is n't one	you will be thinking about it afterwards, i promise you

Conclusion

- Propose model capable of learning **interpretable latent representations** and generating sentences **with specific attributes**
- Variational Auto Encoder + attribute discriminators + independency constraints

Useful Materials:

日本語解説スライド：

https://www.slideshare.net/torufujino/controllable-text-generation-icml-2017-under-review

SGVB:

http://musyoku.github.io/2016/04/29/auto-encoding-variational-bayes/

http://deeplearning.jp/wp-content/uploads/2014/04/20150717-suzuki.pdf

Variational Auto Encoder:

https://www.slideshare.net/ssusere55c63/variational-autoencoder-64515581

S-VAE:

http://musyoku.github.io/2016/07/02/semi-supervised-learning-with-deep-generative-models/

https://www.slideshare.net/beam2d/semisupervised-learning-with-deep-generative-models