VU VRIJE UNIVERSITEIT AMSTERDAM

UNIVERSITY OF AMSTERDAM

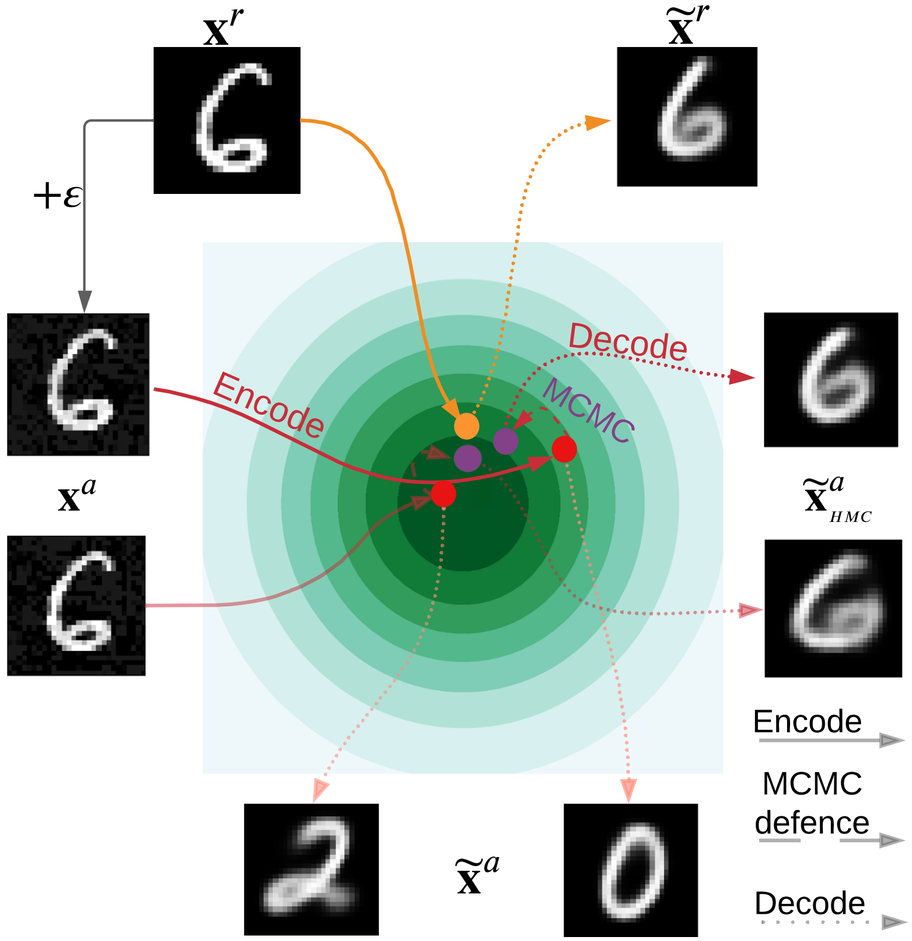
Hybrid Intelligence

Alleviating Adversarial Attacks on Variational Autoencoders with MCMC

Anna Kuzina, Max Welling, Jakub Tomczak

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Adversarial Attack on VAEs



$$x^a = x^r + \varepsilon, \quad \|\varepsilon\| < \delta$$

Unsupervised

x^a "looks" like reference, but is "perceived" differently

$$\varepsilon = \arg\max_{\|\varepsilon\|_p < \delta} \Delta\left[f(x^r + \varepsilon), f(x^r)\right]$$

A. Kuzina, M. Welling, and J. M. Tomczak, "Diagnosing Vulnerability of Variational Auto-Encoders to Adversarial Attacks," in *ICLR 2021 Workshop on Robust and Reliable Machine Learning in the Real World*, 2021.

We focus on unsupervised attacks
defined in the prior works

$$x^a = x^r + \varepsilon, \quad \|\varepsilon\| < \delta$$

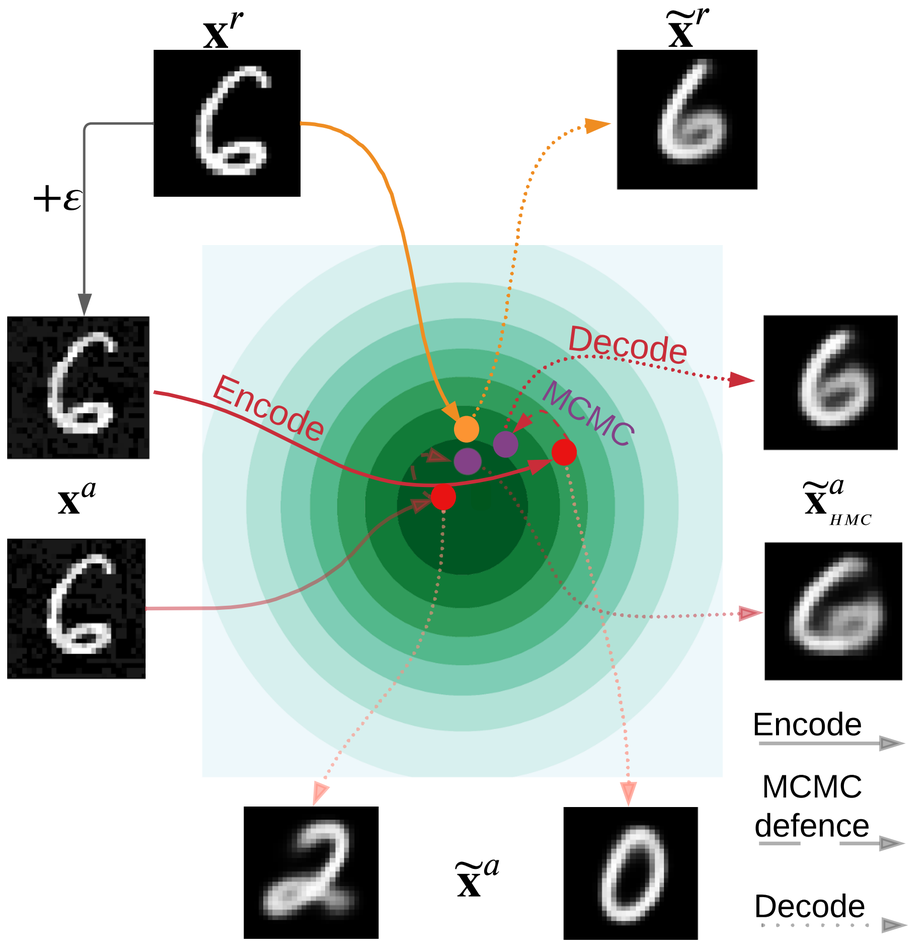
$$\varepsilon = \arg \max_{\|\varepsilon\|_p < \delta} \Delta \left[f(x^r + \varepsilon), f(x^r) \right]$$

Table 1: Different types of attacks on the VAE. We denote $g_\theta(z)$ the deterministic mapping induced by decoder $p_\theta(x|z)$ and as $p_\psi(y|z)$ classification model in the latent space (downstream task).
* Only used during VAE training

	REFERENCE	$f(x)$	$\Delta[A, B]$	$\|\cdot\|_p$	TYPE
Latent Space Attack	(Gondim-Ribeiro et al., 2018; Willetts et al., 2021; Barrett et al., 2021)	$q_\phi(\cdot\|x)$	$\mathrm{KL}[A\|B]$	2	Supervised
Unsupervised Encoder Attack	(Kuzina et al., 2021)	$q_\phi(\cdot\|x)$	$\mathrm{SKL}[A\|B]$	2	Unsupervised
Targeted Output Attack	(Gondim-Ribeiro et al., 2018)	$g_\theta(\tilde{z}), \tilde{z} \sim q_\phi(\cdot\|x)$	$\|A - B\|_2$	2	Supervised
Maximum Damage Attack	(Barrett et al., 2021; Camuto et al., 2021)	$g_\theta(\tilde{z}), \tilde{z} \sim q_\phi(\cdot\|x)$	$\|A - B\|_2$	2	Unsupervised
Projected Gradient Descent Attack*	(Cemgil et al., 2019)	$q_\phi(\cdot\|x)$	$\mathcal{WD}[A, B]$	inf	Unsupervised
Adversarial Accuracy	(Cemgil et al., 2019; 2020)	$p_\psi(y\|\tilde{z}), \tilde{z} \sim q_\phi(\cdot\|x)$	CROSS ENTROPY	inf	Unsupervised

Defence Strategy

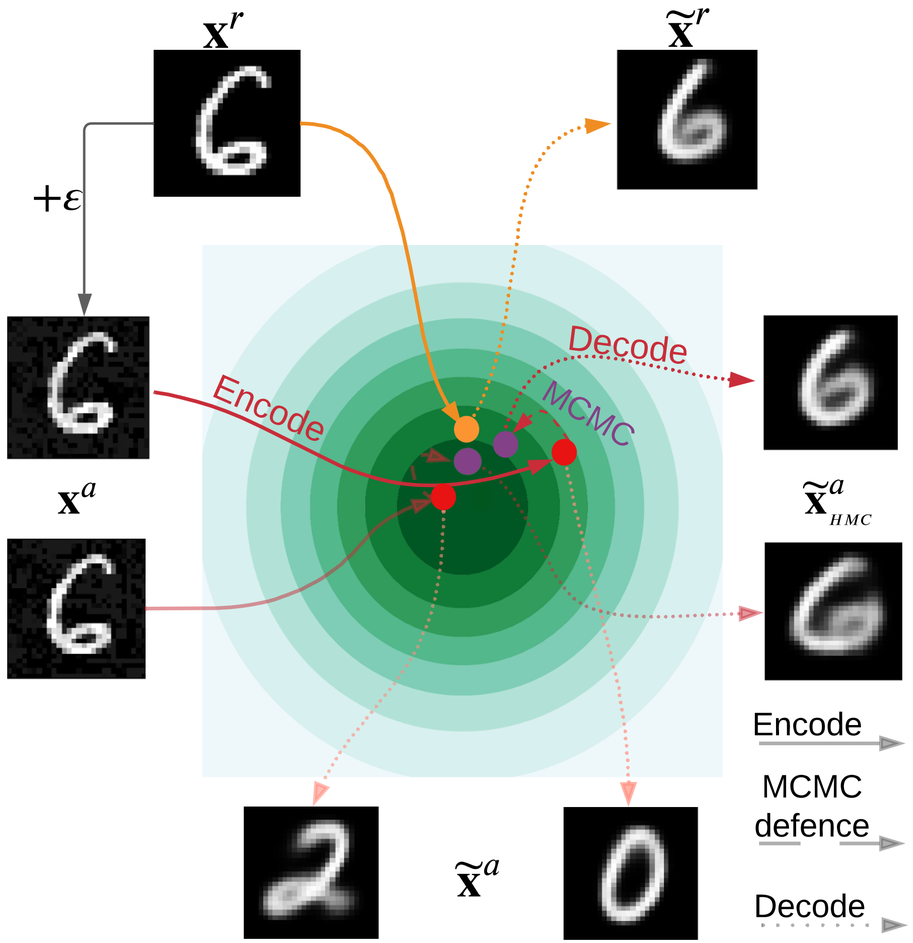
$$x^a = x^r + \varepsilon, \quad \|\varepsilon\| < \delta$$



$z^a \sim q_\phi(z|x^a)$ **vs** $z^r \sim q_\phi(z|x^r)$

Defence Strategy

$$x^a = x^r + \varepsilon, \quad \|\varepsilon\| < \delta$$



$z^a \sim q_\phi(z \mid x^a)$ **vs** $z^r \sim q_\phi(z \mid x^r)$

Let's use samples from the true posterior instead:

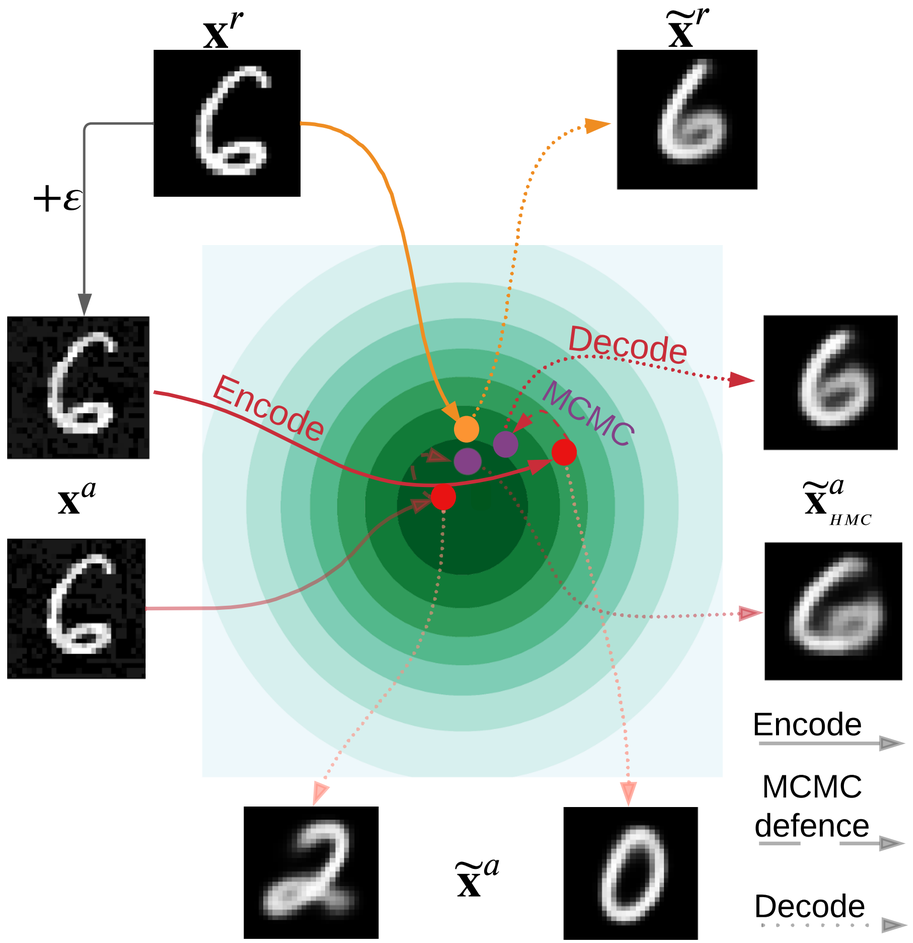
For that we use t steps of MCMC (starting from the encoder):

$$z^{(t)} \sim q^{(t)}(z \mid x^a) = \int q_\phi(z_0 \mid x^a) Q^{(t)}(z \mid z_0) dz_0$$

with target density $p_\theta(z \mid x^a) \propto p(z) p_\theta(x^a \mid z)$

Defence Strategy

$$x^a = x^r + \varepsilon, \quad \|\varepsilon\| < \delta$$



$z^a \sim q_\phi(z \mid x^a)$ **vs** $z^r \sim q_\phi(z \mid x^r)$

Let's use samples from the true posterior instead:

For that we use t steps of MCMC (starting from the encoder):

$$z^{(t)} \sim q^{(t)}(z \mid x^a) = \int q_\phi(z_0 \mid x^a) Q^{(t)}(z \mid z_0) dz_0$$

with target density $p_\theta(z \mid x^a) \propto p(z) p_\theta(x^a \mid z)$

Each step brings us "closer" to the true posterior:

$$\mathrm{KL}\left[q^{(t)}(z \mid x^a) \,||\, p_\theta(z \mid x^a)\right] \leq \mathrm{KL}\left[q^{(t-1)}(z \mid x^a) \,||\, p_\theta(z \mid x^a)\right]$$

Final Algorithm

1. *(Defender)*

Train a VAE:

$$q_\phi(z|x), p(z), p_\theta(x|z)$$

2. *(Attacker)*

For a given x^r, construct the attack x^a

$$x^a = x^r + \varepsilon, \quad \|\varepsilon\| < \delta$$

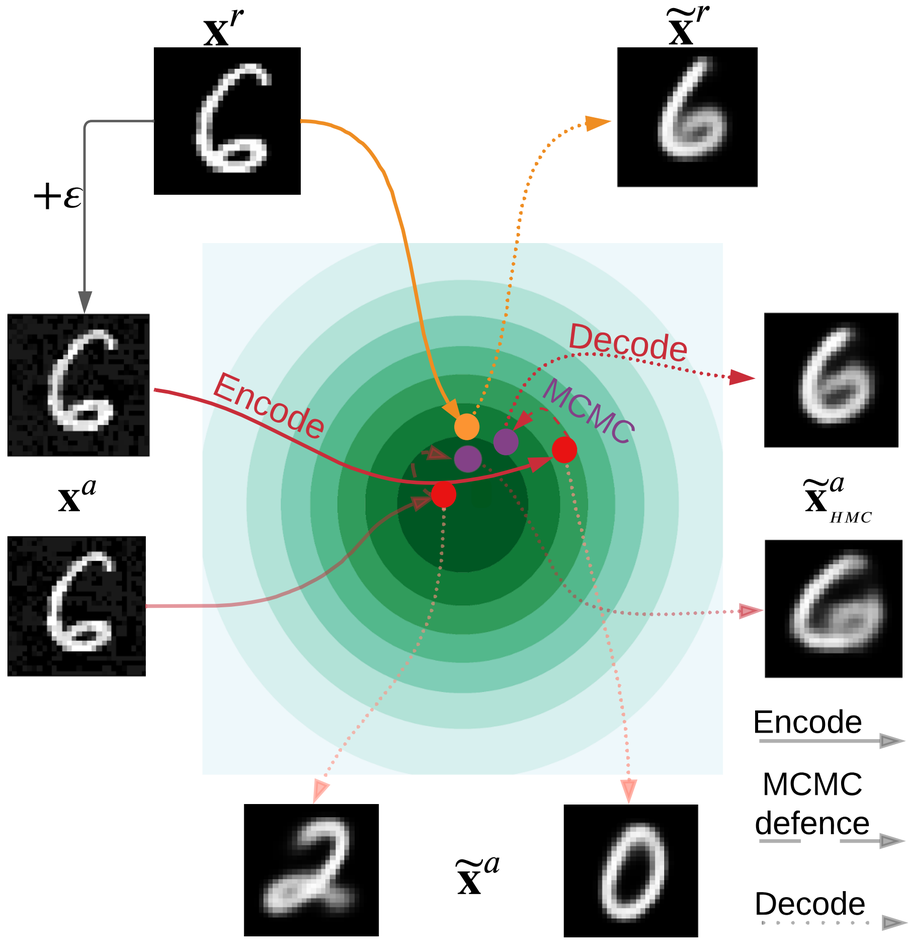
s.t $q_\phi(z|x^a)$ is "far enough" from $q_\phi(z|x^r)$

3. *(Defender)*

Initialize the latent code $z_0 \sim q_\phi(z|x^a)$

Run T steps of HMC with the target $\propto p(z)p_\theta(x^a|z)$

Use $z := z^{(T)}$ to decode / in downstream task



* Note that q_ϕ and p_θ can be of any form, e.g., hierarchical VAEs

Why it works?

Theoretical justification

$$x^a = x^r + \varepsilon, \quad \|\varepsilon\| < \delta$$

$z^r \sim q_\phi(z \mid x^r)$ **vs** $z^{(t)} \sim q^{(t)}(z \mid x^a)$

Why it works?

Theoretical justification

$$x^a = x^r + \varepsilon, \quad \|\varepsilon\| < \delta$$

$$z^r \sim q_\phi(z \mid x^r) \quad \textbf{vs} \quad z^{(t)} \sim q^{(t)}(z \mid x^a)$$

Theorem:

$$\mathrm{TV}[q^{(t)}(z \mid x^a) \| q_\phi(z \mid x^r)] \leq \sqrt{\tfrac{1}{2}\mathrm{KL}\left[q^{(t)}(z \mid x^a) \| p_\theta(z \mid x^a)\right]} + \sqrt{\tfrac{1}{2}\mathrm{KL}\left[q_\phi(z \mid x^r) \| p_\theta(z \mid x^r)\right]} + o(\sqrt{\|\varepsilon\|})$$

How good is defence

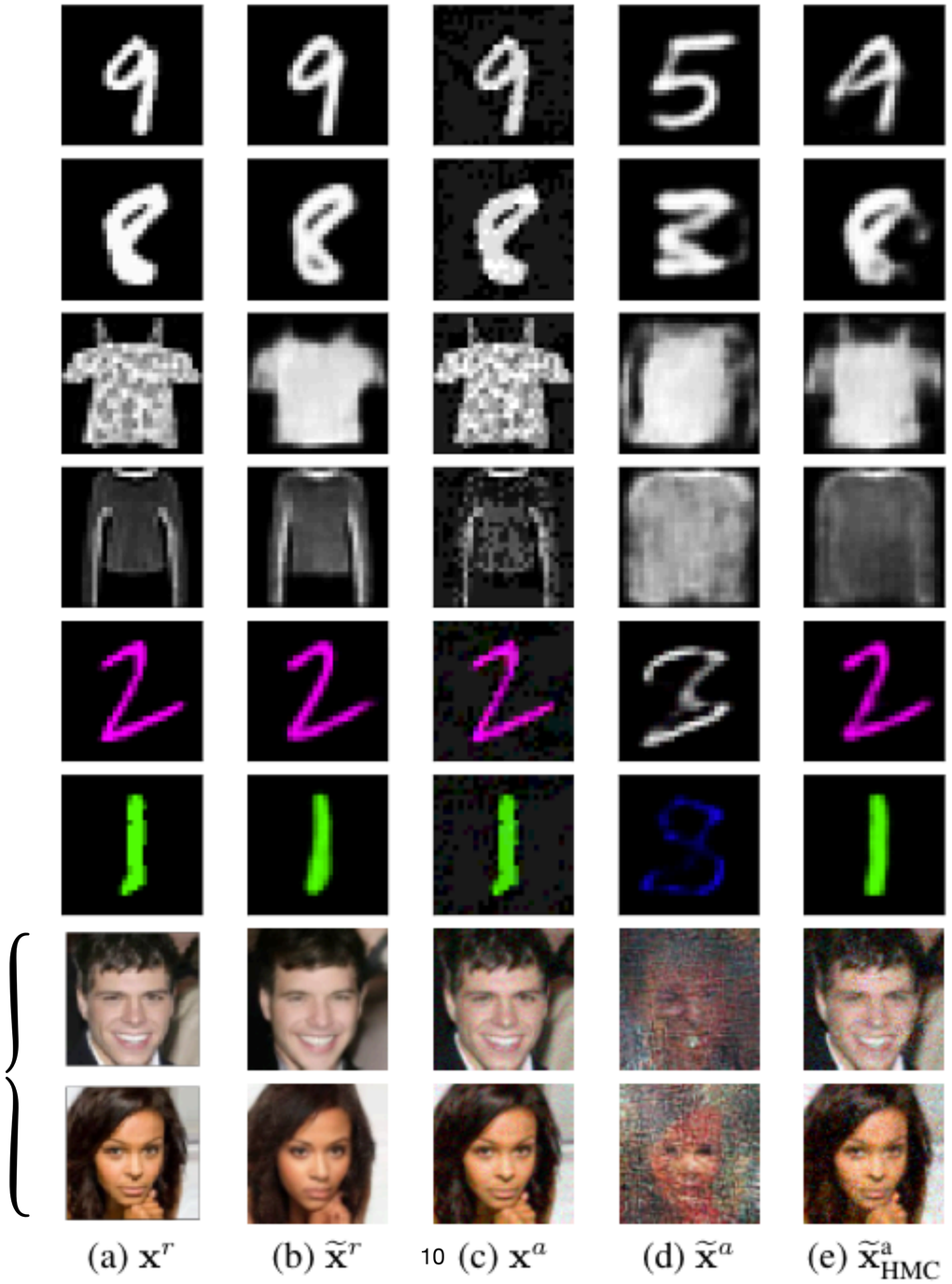
Goes to 0 with t

How good VAE is (*approximation gap*)

Attack radius

Results

NVAE:
deep hierarchical VAE



(a) $\mathbf{x}^r$ (b) $\widetilde{\mathbf{x}}^r$ (c) $\mathbf{x}^a$ (d) $\widetilde{\mathbf{x}}^a$ (e) $\widetilde{\mathbf{x}}^{\mathrm{a}}_{\mathrm{HMC}}$

Thank you

Paper:

Code:

"Alleviating Adversarial Attacks on Variational Autoencoders with MCMC", Neural Information Processing Systems (NeurIPS 2022).
Anna Kuzina, Max Welling, Jakub Tomczak

a.kuzina@vu.nl

Why it works?

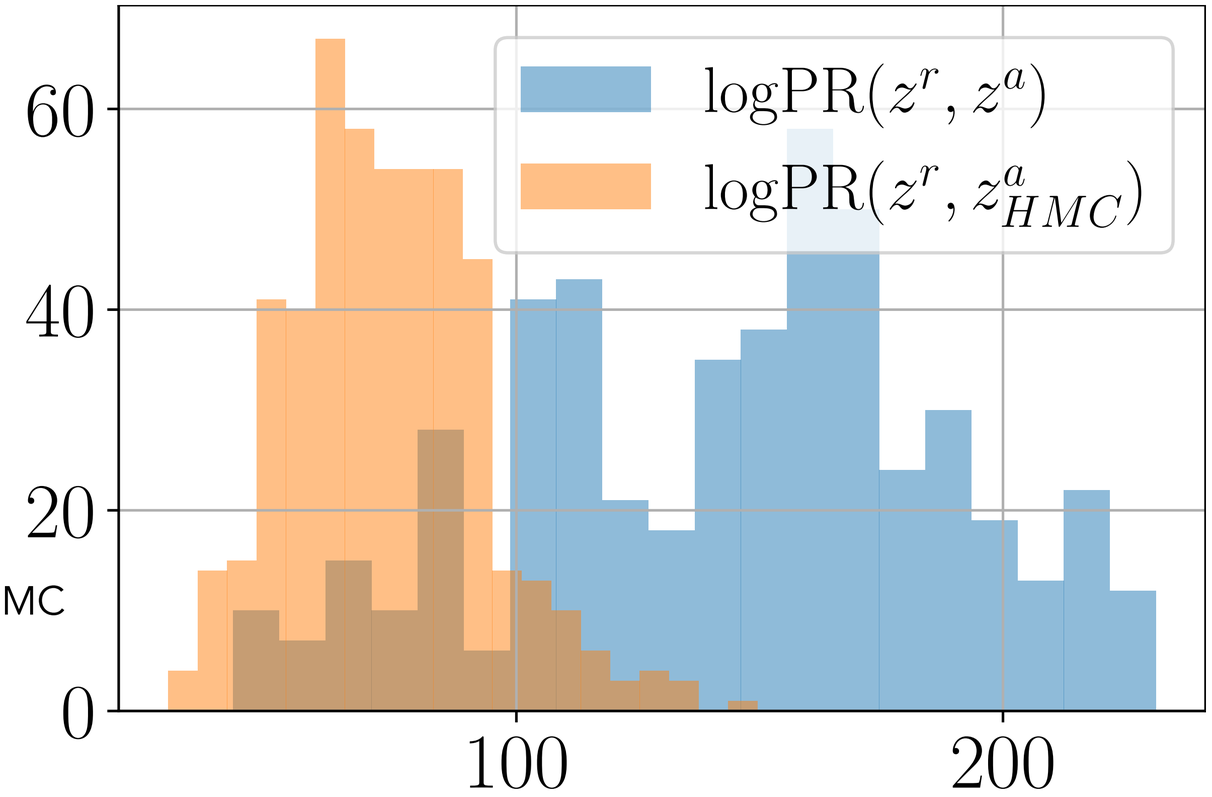
Empirical Evidence

Given a reference point, one can evaluate posterior ratio for two latent codes:

$$\mathrm{PR}(z_1, z_2) = \frac{p_\theta(z_1 \,|\, x^r)}{p_\theta(z_2 \,|\, x^r)}$$

Blue: reference latent code VS adversarial latent code

Orange: reference latent code VS adversarial latent code after HMC



What if attacker knowns the defence strategy?