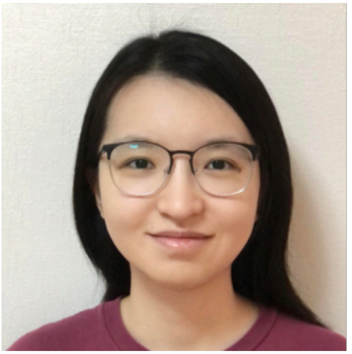
NEURAL INFORMATION PROCESSING SYSTEMS

Google DeepMind

THE UNIVERSITY OF BRITISH COLUMBIA

Adaptive Randomized Smoothing: Certified Adversarial Robustness for Multi-Step Defences

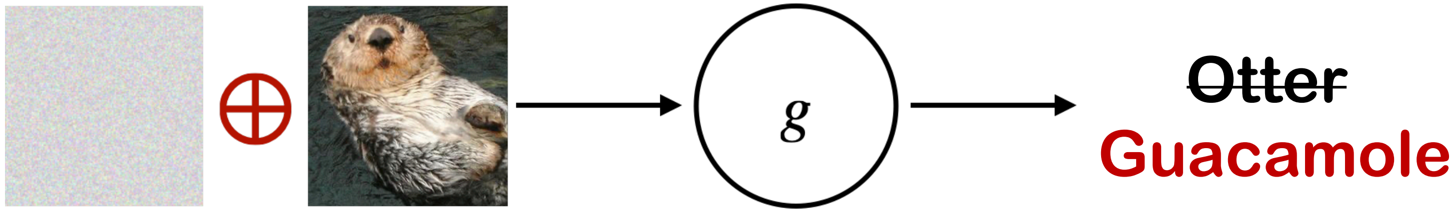
Saiyue Lyu*, Shadab Shaikh*, Frederick Shpilevskiy*, Evan Shelhamer, Mathias Lécuyer





Adversarial Examples

- **Adversarial Examples (AE): test-time attacks to control model predictions with small crafted input perturbations.**

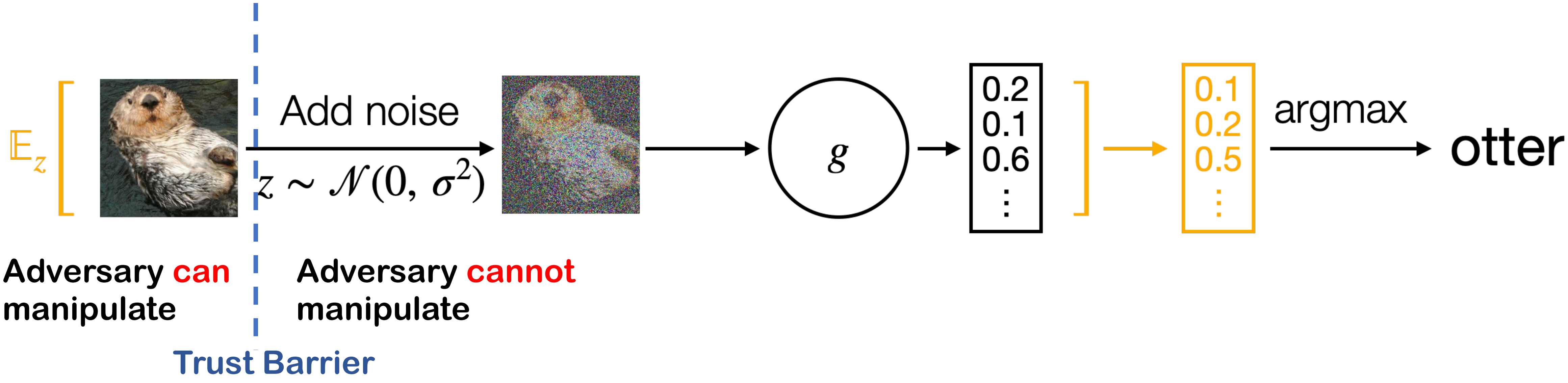


- **The power of the adversary is determined by the maximum size of the attack:**

ℓ_2 attack: $||$  $||_2$ ℓ_∞ attack: $||$  $||_\infty$

- **Randomized Smoothing can provide provable defenses against adversarial examples!**

Randomized Smoothing (RS)



- **Randomized Smoothing gives provable robustness by averaging over noisy predictions:**
- **Theorem (Cohen et al. 2019): with** $\mathbb{P}(f(X+z)=y_+) \geq \underline{p_+} \geq \overline{p_-} \geq \max_{y_- \neq y_+} \mathbb{P}(f(X+z)=y_-)$ **we have:**

No ℓ_2 attack with $\| \cdot \|_2 \leq r_X = \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p_+}) - \Phi^{-1}(\overline{p_-})\right)$ certificate

Source: Cohen et. al., "Certified adversarial robustness via randomized smoothing", ICML 2019.

Limitations of RS

- Noise degrades accuracy.
- Difficulty scaling to high dimensional inputs for ℓ_∞ threat models:

$r_X \leq \|$  $\|_2 \leq \sqrt{d}\|$  $\|_\infty$

- Does not support test-time adaptivity to adapt the accuracy/robustness trade-off to the input.

We use Gaussian differential privacy to address these shortcomings!

Gaussian Differential Privacy

- **We can frame privacy as a hypothesis test between $\mathcal{H}_0 : D$ and $\mathcal{H}_1 : D'$ (i.e. does $x \in D$?). This enables a hypothesis test definition of DP.**
- **A tradeoff function f bounds the power of any statistical test of $\mathcal{H}_0$ v.s. $\mathcal{H}_1$.**

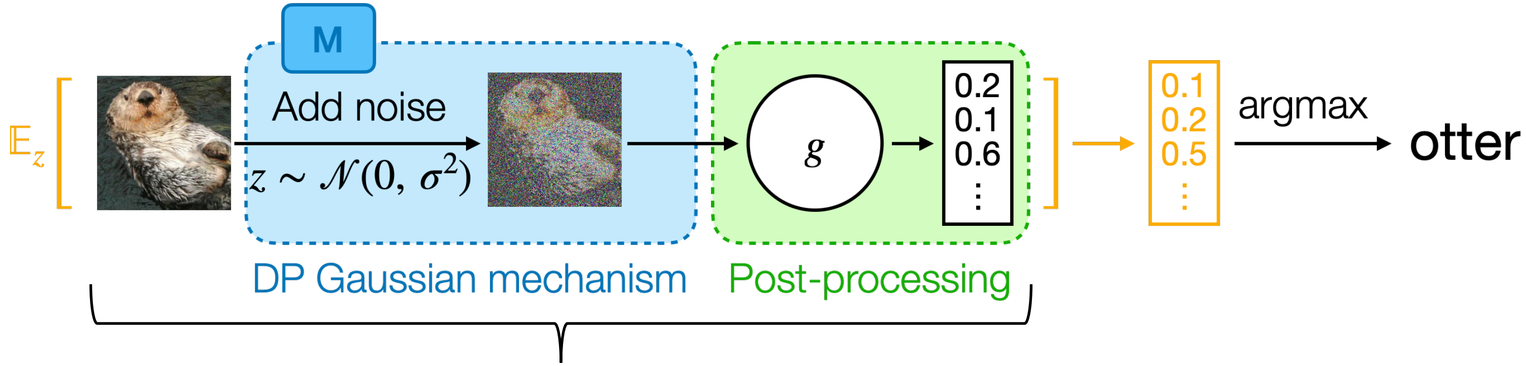
(Theorem 2.7 Dong et al. 2019) For a Gaussian mechanism $\mathcal{M}(D) = \theta(D) + z, z \sim \mathcal{N}\left(0, \frac{r^2}{\mu^2}\right)$, such that for any neighboring $D, D', \theta(D)$- $\theta(D') \in B_2(r)$ (i.e., the ℓ_2 sensitivity of θ is r), we have that $\mathcal{M}$ is G_μ-DP with function $f = G_\mu$ defined by :

$$G_\mu(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \mu), \text{ for all } \alpha \in [0,1]$$

- **Composition:** the composition of an G_{μ_1}-DP Gaussian mechanism and an G_{μ_2}-DP Gaussian mechanism is G_μ-DP Gaussian mechanism with $\mu = \sqrt{\mu_1^2 + \mu_2^2}$.

Source: Jinshuo Dong, Aaron Roth, and Weijie J Su. 'Gaussian differential privacy'. In: arXiv (2019), Journal of the Royal Statistical Society (2022).

GDP and Randomized Smoothing



GDP under neighbouring definition $D' = D + \delta,\ \|\delta\|_p \leq r.$

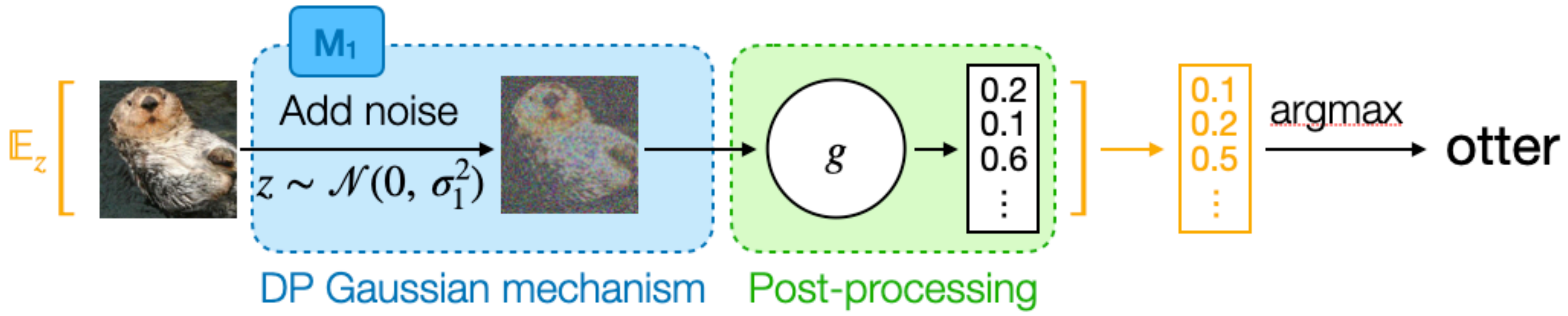
- **We prove that our GDP randomized smoothing mechanism satisfies**

$$f(1-\underline{p_+}) \geq 1 - f(\overline{p_-}) \Rightarrow \forall \|\delta\|_p \leq r,\ M_S(D+\delta) = y_+$$

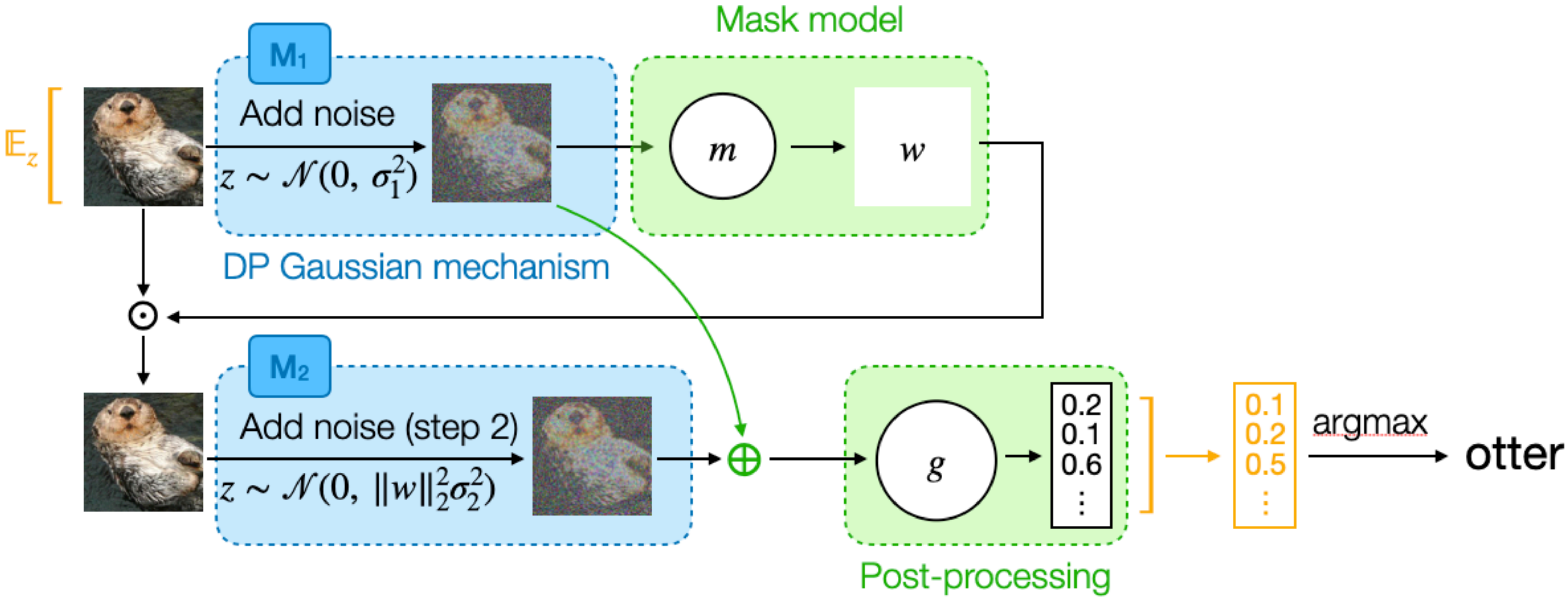
- **Using this result and GDP, we prove that:**

No ℓ_2 attack is possible with || $\|_2 \leq r_X = \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p_+}) - \Phi^{-1}(\overline{p_-})\right)$

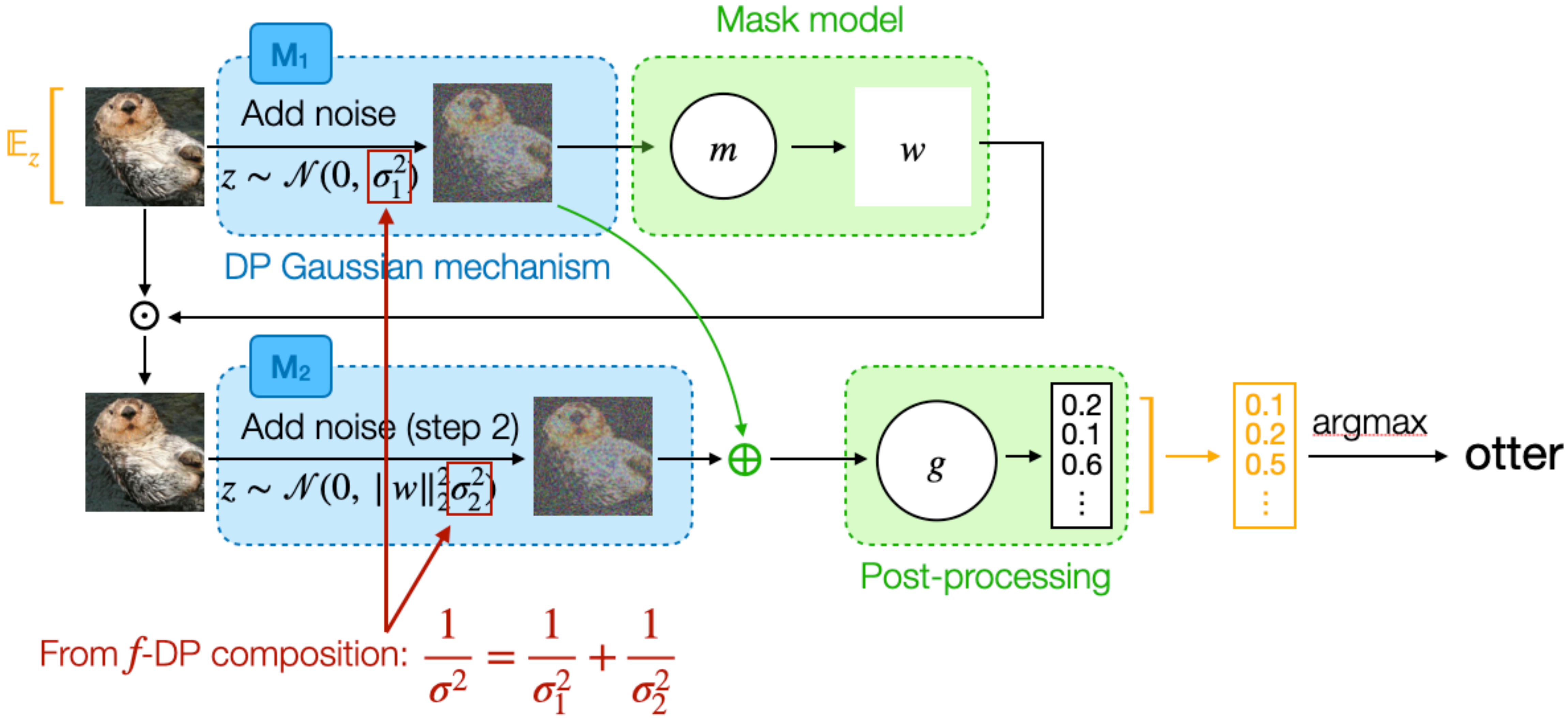
Adaptive Randomized Smoothing for ℓ_∞



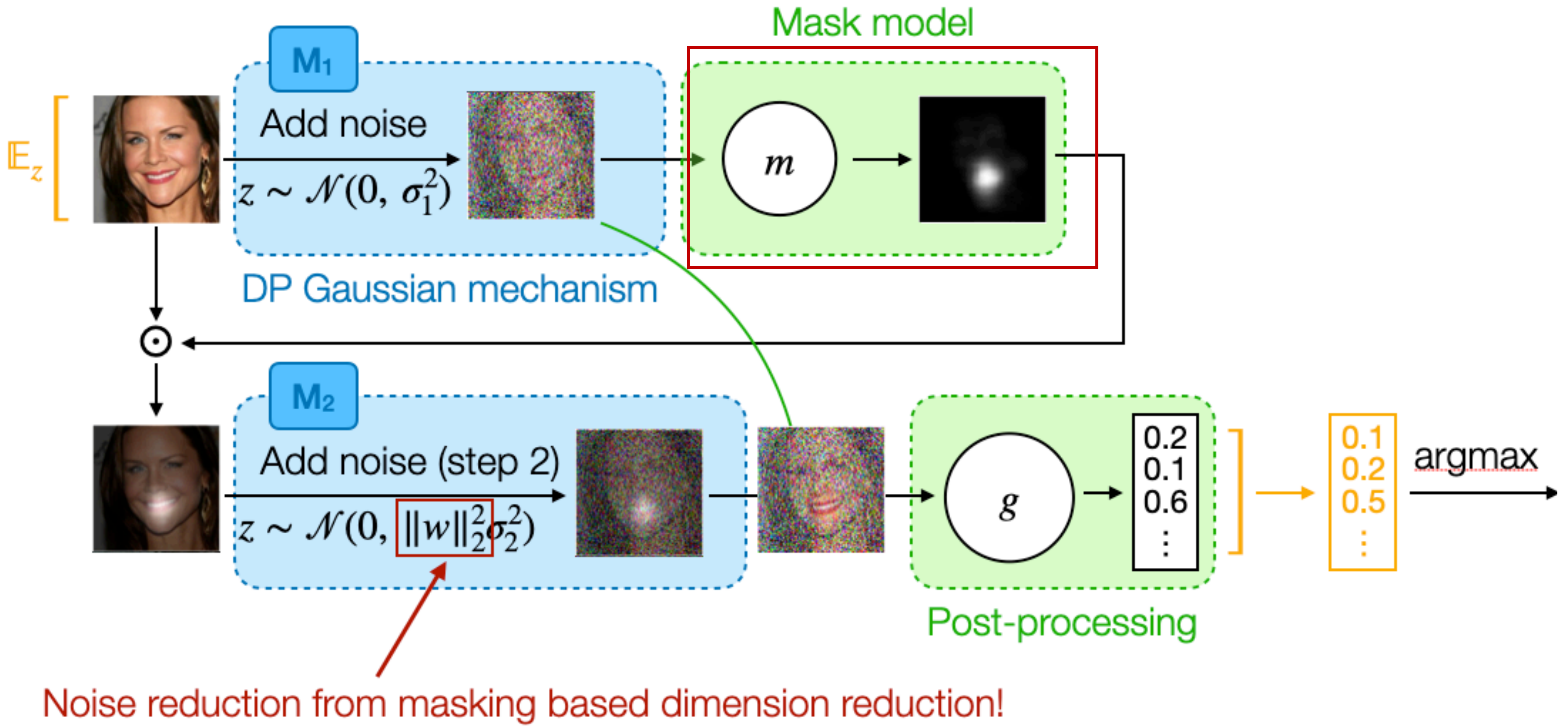
Adaptive Randomized Smoothing for ℓ_∞



Adaptive Randomized Smoothing for ℓ_∞



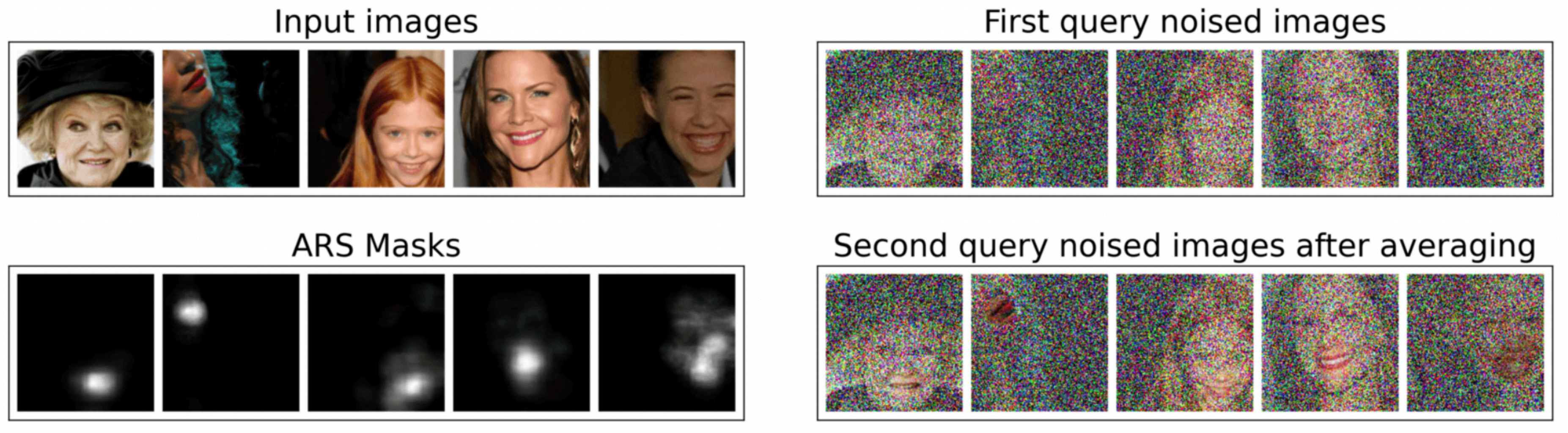
Why does ARS help?

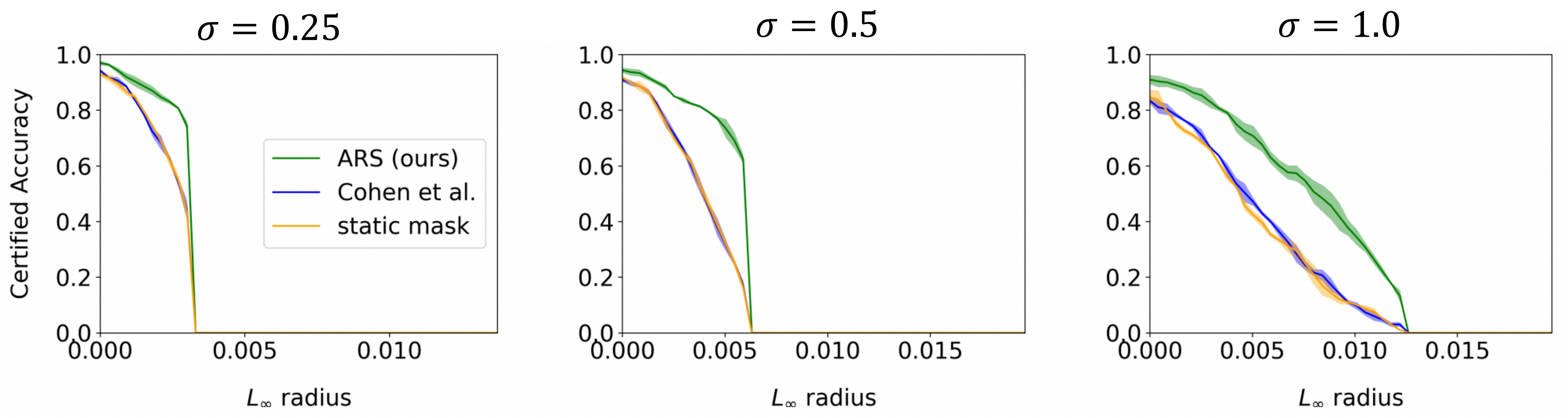


mouth slightly open

Evaluations

- **CelebA**





Conclusion

- Adaptive Randomized Smoothing (ARS) uses DP composition post-processing properties to certify adaptive multi-step models.
- ARS learns to adjust the scale of noise based on the test input.
- ARS provides higher accuracy at a given level of provable robustness.

Link to our code


THANK
YOU