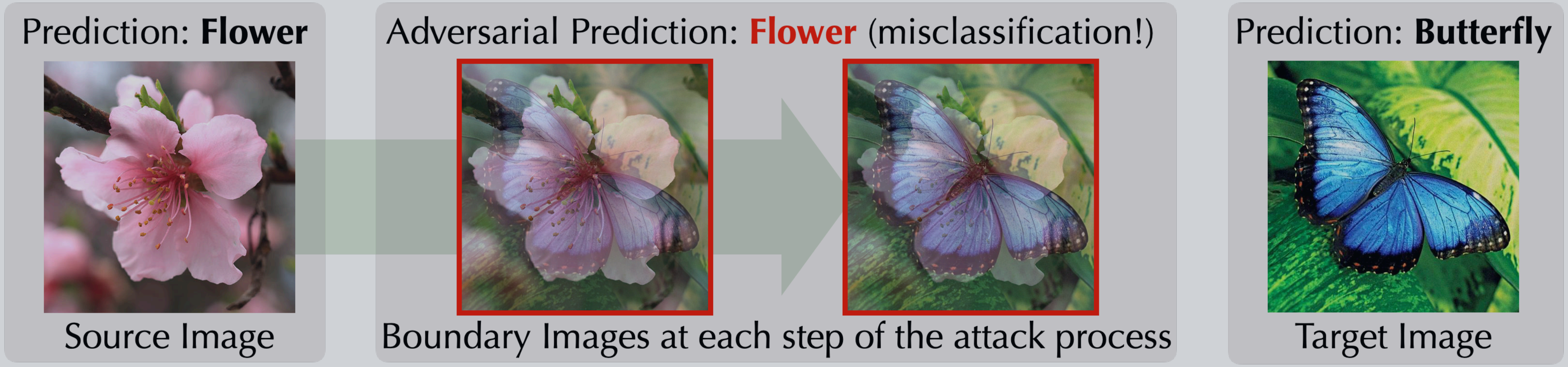
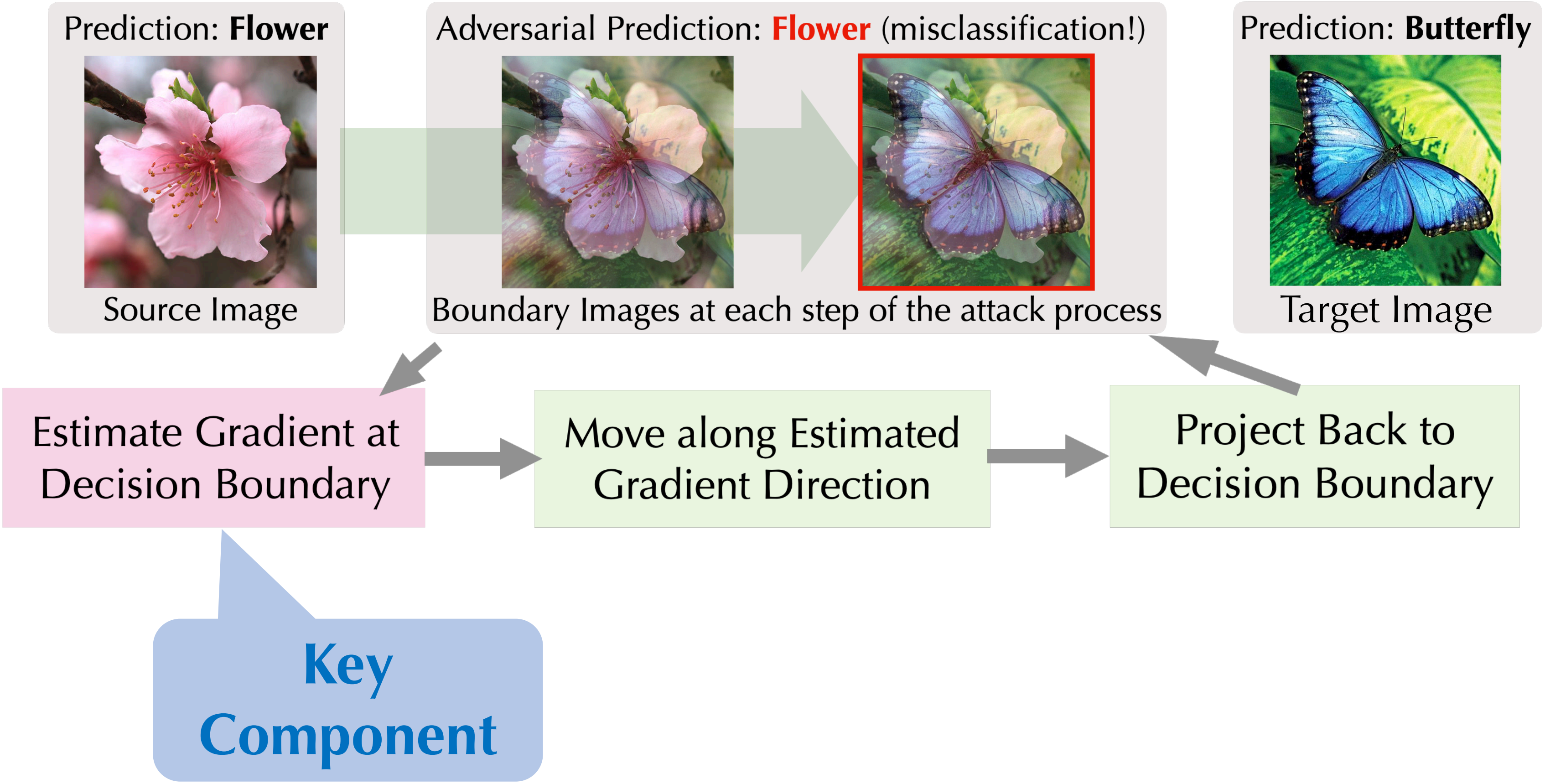



# Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation

Jiawei Zhang*, **Linyi Li***, Huichen Li,
Xiaolu Zhang, Shuang Yang, Bo Li

# Background: Boundary Blackbox Attack

- **Boundary Blackbox Attack**: an effective attack for neural network
  - Require **minimum** query information (decision) of the target model

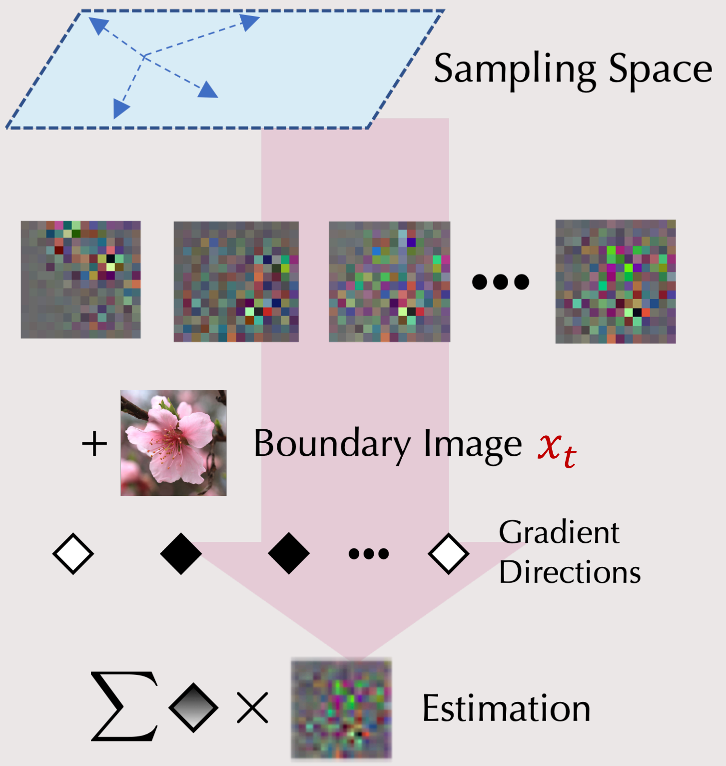


- How to reduce dimensionality for gradient estimation?
- What's the optimal projection scale for estimating the gradient?
- How to select the "optimal" projection scale in practice?

# How to Estimate Boundary Gradient?

A general framework: sampling based approach combined with projection

| | |
|---|---|
| Sample Vectors from Low-Dimensional Sampling Space | $\{u_b\}_{b=1}^{B}$ are $m$-dimentional unit vectors |
| Map to **Progressive-Scaled Projection Subspace** via Projection Function $\boldsymbol{f}$ | $\boldsymbol{f}\colon \mathbb{R}^m$(sampling space) $\rightarrow \mathbb{R}^n$(original space) <br> $\{u_b\}_{b=1}^{B} \mapsto \{\underbrace{\boldsymbol{f}(\delta u_b) - \boldsymbol{f}(0)}_{:=\Delta \boldsymbol{f}(\delta u_b)}\}_{b=1}^{B}$ |
| Query Target Model | $\{\mathrm{sgn}(G(x_t + \Delta \boldsymbol{f}(\delta u_b)))\}_{b=1}^{B}$ <br> ($G$: target model) |
| Estimate Boundary Gradient | $\widetilde{\nabla G}(x_t) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{sgn}(G(x_t + \Delta \boldsymbol{f}(\delta u_b)))\Delta \boldsymbol{f}(\delta u_b)$ |



# A Systematic Analysis of Gradient Estimator

We provide:

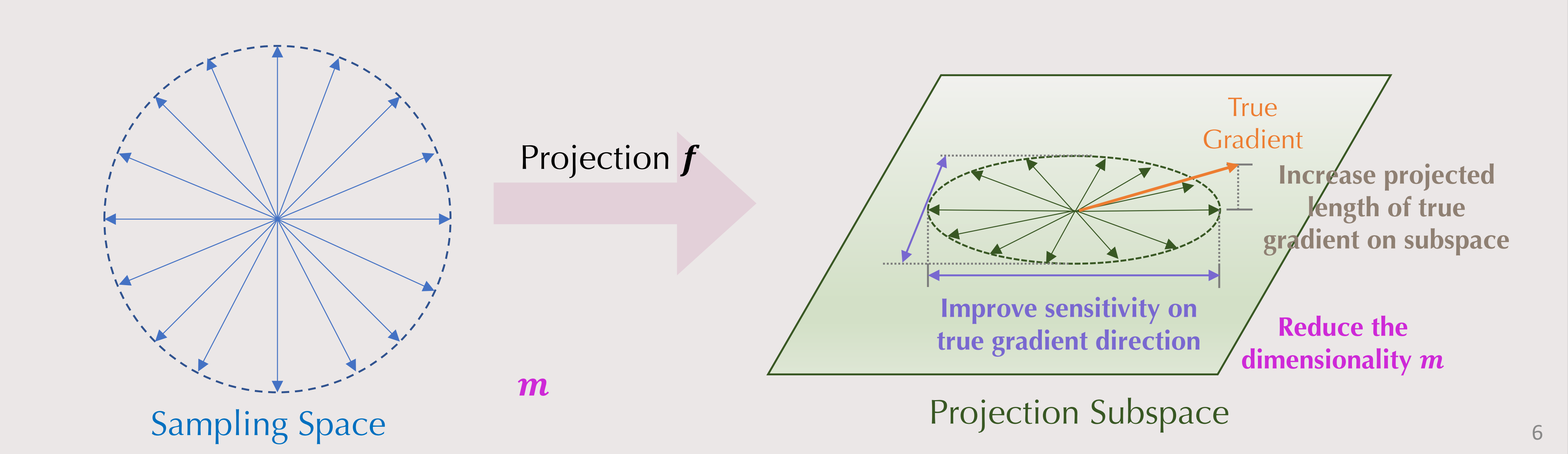
- **Tighter** and **more general** expectation lower bound
- **First concentration** lower bound

for cosine similarity between estimated and true gradient

# Key Characteristics

- What contributes to query-efficient & accurate gradient estimation?

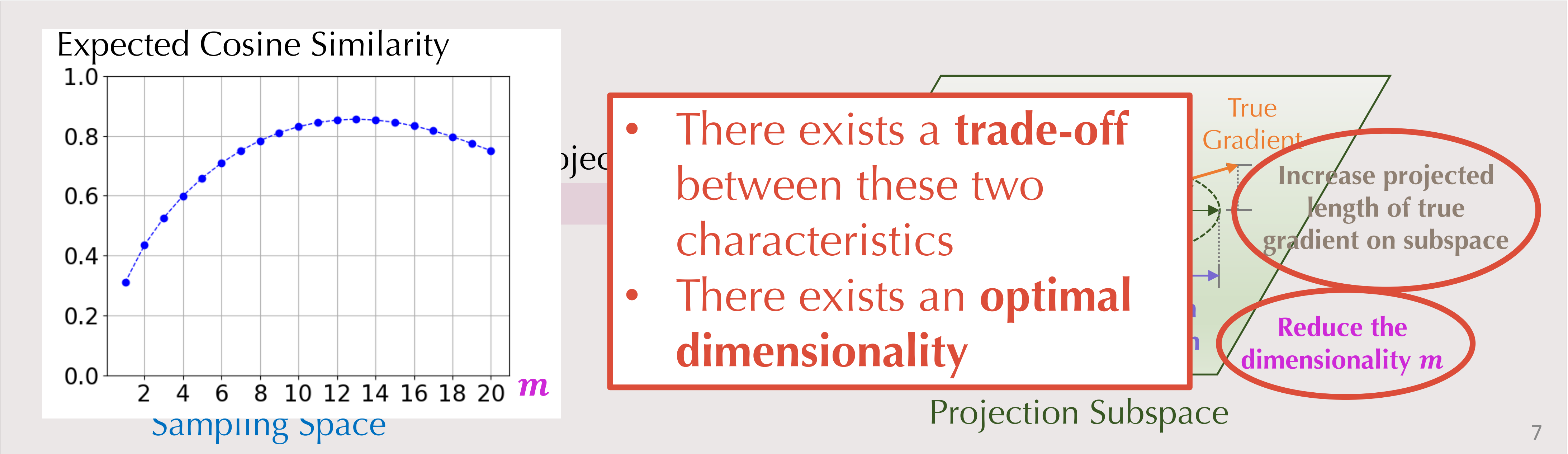
$$\cos\langle\widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \geq \frac{\|\mathrm{proj}_{\nabla f(0)} \nabla S(x_t)\|_2}{\|\nabla S(x_t)\|_2} \cdot \left(1 - \mathcal{O}\left(m^2 \cdot \frac{\sum_{i=2}^{m} \alpha_i^2}{m-1}\left(\overbrace{\frac{\delta^2\beta_f^2}{\alpha_1^4} + \frac{\alpha_{\max}^4}{\alpha_1^4} \cdot \frac{\delta^2\beta_S^2}{\|\mathrm{proj}_{\nabla f(0)} \nabla S(x_t)\|_2^2}}^{\text{expectation}} + \overbrace{\frac{\ln(m/p)}{B\alpha_1^2}}^{\text{sampling error}}\right)\right)\right)$$



# Key Characteristics

- What contributes to query-efficient & accurate gradient estimation?

$$\cos\langle \widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \geq \frac{\|\mathrm{proj}_{\nabla f(0)} \nabla S(x_t)\|_2}{\|\nabla S(x_t)\|_2}\cdot\left(1-\mathcal{O}\left(m^2\cdot\frac{\sum_{i=2}^{m} \alpha_i^2}{m-1}\left(\overbrace{\frac{\delta^2\beta_f^2}{\alpha_1^4}+\frac{\alpha_{\max}^4}{\alpha_1^4}\cdot\frac{\delta^2\beta_S^2}{\|\mathrm{proj}_{\nabla f(0)}\nabla S(x_t)\|_2^2}}^{\text{expectation}}+\overbrace{\frac{\ln(m/p)}{B\alpha_1^2}}^{\text{sampling error}}\right)\right)\right)$$



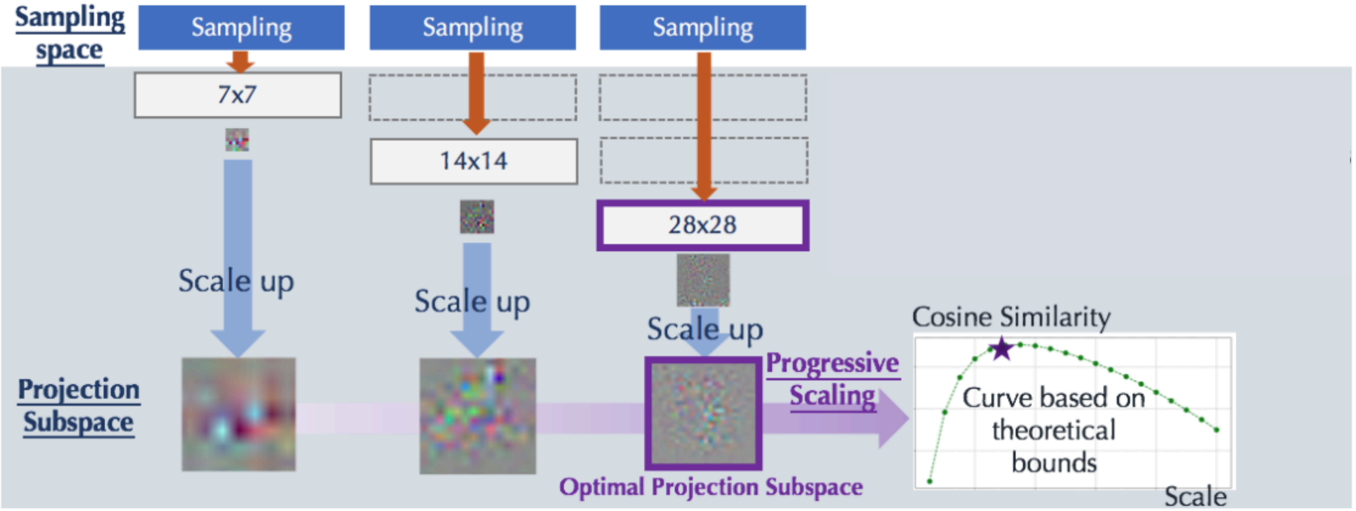

- There exists a **trade-off** between these two characteristics
- There exists an **optimal dimensionality**

# Progressive-Scale enabled projective Boundary blackbox Attack (PSBA)

**Focusing on low-frequency subspace ⇔ Perturbing at some small resolution (scale)**

- Use Progressive-GAN as the projection model
  - Training ⇔ Learning true gradient information at corresponding scales

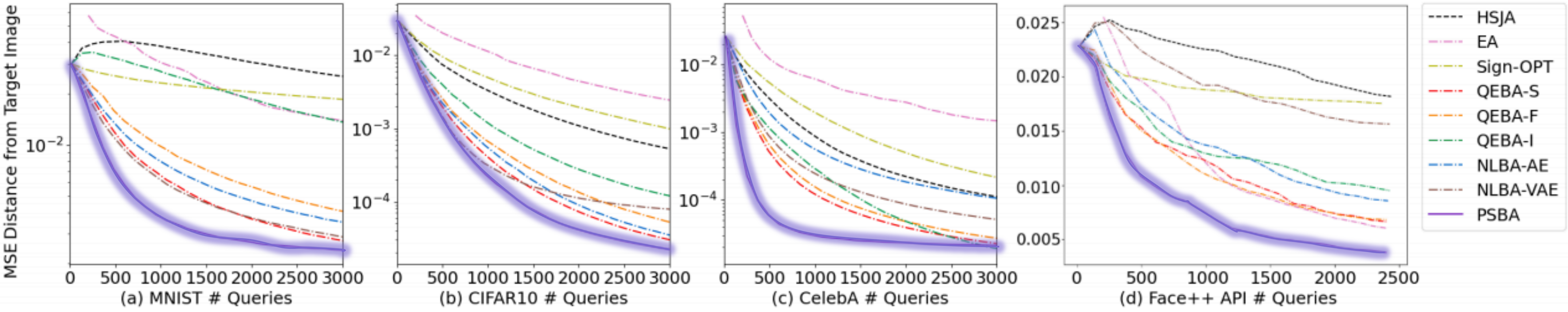
- Trade-off exists = optimal scale exists



# PSBA Performance

With more query-efficient gradient estimation, PSBA **significantly** outperforms baselines

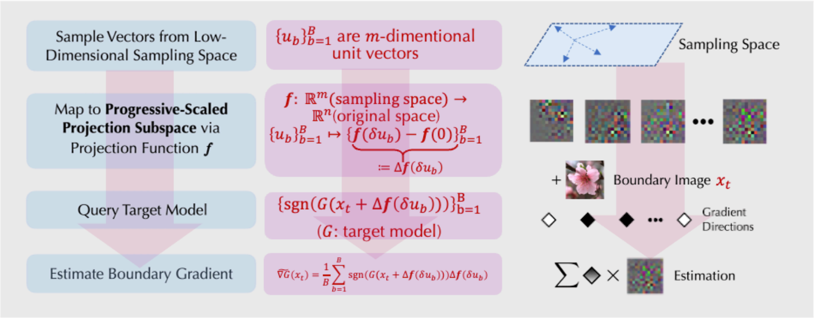
- Finds adversarial examples with much smaller $\ell_2$ distance under small query budget

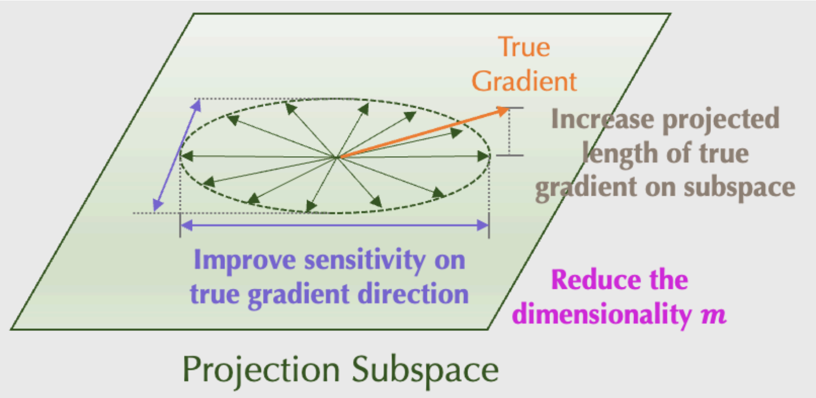


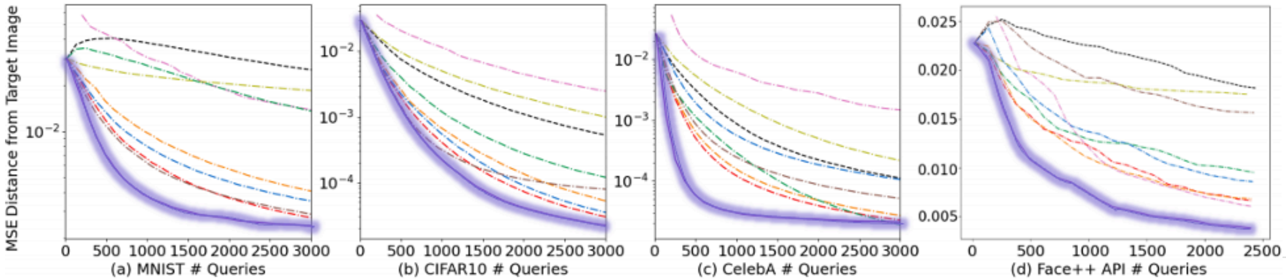
$$\cos\langle \widetilde{\nabla S}(x_t), \nabla S(x_t)\rangle \ge \frac{\|\mathrm{proj}_{\nabla f(0)} \nabla S(x_t)\|_2}{\|\nabla S(x_t)\|_2} \cdot \left(1 - \mathcal{O}\left(m^2 \cdot \frac{\sum_{i=2}^m \alpha_i^2}{m-1}\left(\overbrace{\frac{\delta^2\beta_f^2}{\alpha_1^4} + \frac{\alpha_{\max}^4}{\alpha_1^4}\cdot\frac{\delta^2\beta_S^2}{\|\mathrm{proj}_{\nabla f(0)}\nabla S(x_t)\|_2^2}}^{\text{expectation}} + \overbrace{\frac{\ln(m/p)}{B\alpha_1^2}}^{\text{sampling error}}\right)\right)\right)$$

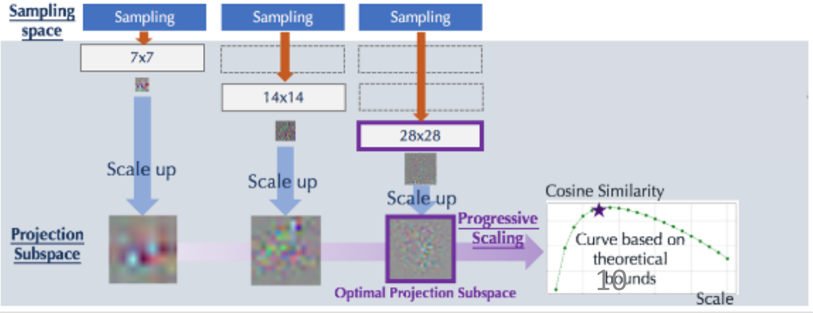
# Summary

- Theoretical framework to analyze gradient estimation in boundary blackbox attacks
- Characterize key characteristics and trade-offs in gradient estimation
- Propose PSBA, a theory motivated and query efficient blackbox attack
- Extensive experimental evaluation on several datasets and a commercial API




AI-secure/PSBA