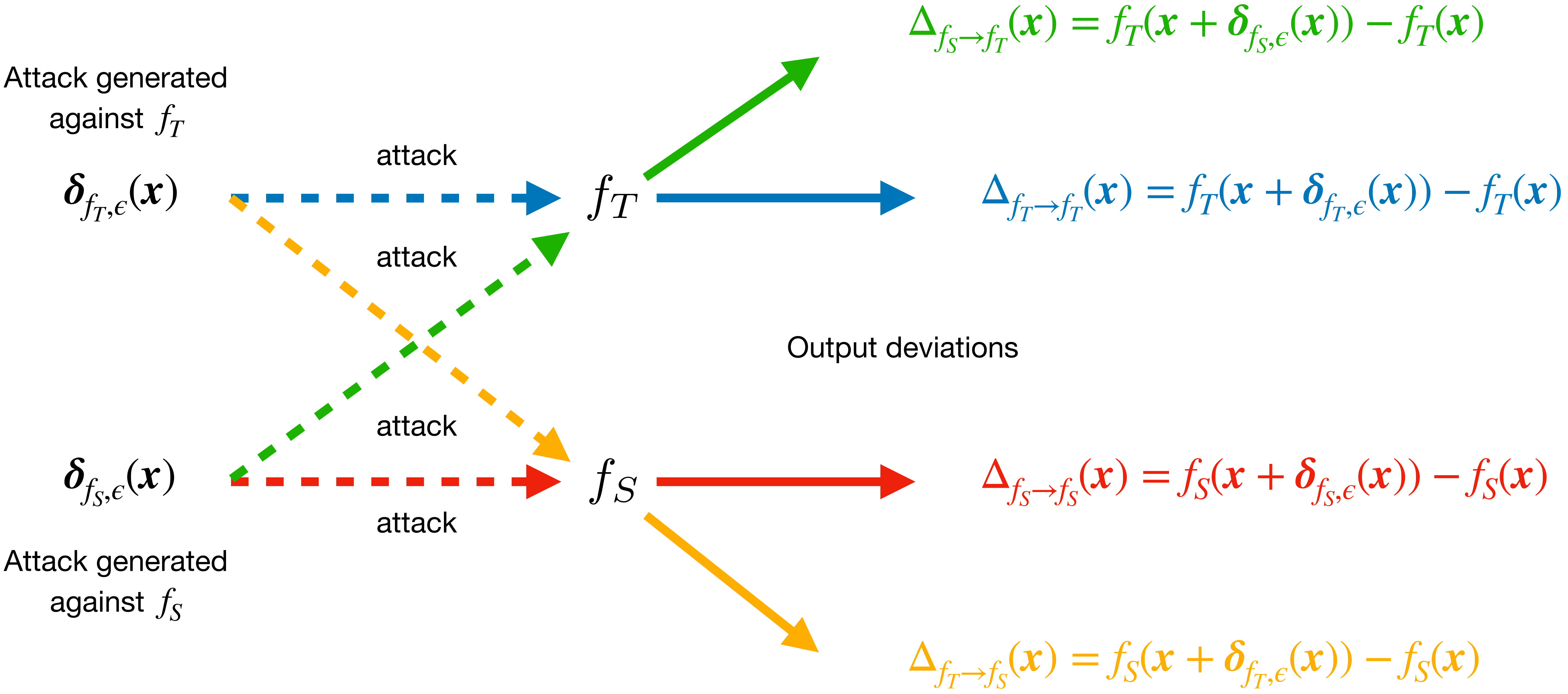
$$\alpha_1^{f_S\to f_T}(\boldsymbol{x}) = \frac{\ell_{adv}(f_T(\boldsymbol{x}), f_T(\boldsymbol{x}+\boldsymbol{\delta}_{f_S,\epsilon}(\boldsymbol{x})))}{\ell_{adv}(f_T(\boldsymbol{x}), f_T(\boldsymbol{x}+\boldsymbol{\delta}_{f_T,\epsilon}(\boldsymbol{x})))}$$

$$\alpha_2^{f_S\to f_T} = \|\mathbb{E}_{\boldsymbol{x}\sim\mathscr{D}}[\widehat{\Delta_{f_S\to f_S}(\boldsymbol{x})}\widehat{\Delta_{f_S\to f_T}(\boldsymbol{x})}^\top]\|_F$$


Attack generated against f_T
$\boldsymbol{\delta}_{f_T,\epsilon}(\boldsymbol{x})$
attack
attack
attack
attack
$\boldsymbol{\delta}_{f_S,\epsilon}(\boldsymbol{x})$
Attack generated against f_S
f_T
f_S
Output deviations

$$\Delta_{f_S\to f_T}(\boldsymbol{x}) = f_T(\boldsymbol{x}+\boldsymbol{\delta}_{f_S,\epsilon}(\boldsymbol{x})) - f_T(\boldsymbol{x})$$

$$\Delta_{f_T\to f_T}(\boldsymbol{x}) = f_T(\boldsymbol{x}+\boldsymbol{\delta}_{f_T,\epsilon}(\boldsymbol{x})) - f_T(\boldsymbol{x})$$

$$\Delta_{f_S\to f_S}(\boldsymbol{x}) = f_S(\boldsymbol{x}+\boldsymbol{\delta}_{f_S,\epsilon}(\boldsymbol{x})) - f_S(\boldsymbol{x})$$

$$\Delta_{f_T\to f_S}(\boldsymbol{x}) = f_S(\boldsymbol{x}+\boldsymbol{\delta}_{f_T,\epsilon}(\boldsymbol{x})) - f_S(\boldsymbol{x})$$

$$\alpha_1^{f_T\to f_S}(\boldsymbol{x}) = \frac{\ell_{adv}(f_S(\boldsymbol{x}), f_S(\boldsymbol{x}+\boldsymbol{\delta}_{f_T,\epsilon}(\boldsymbol{x})))}{\ell_{adv}(f_S(\boldsymbol{x}), f_S(\boldsymbol{x}+\boldsymbol{\delta}_{f_S,\epsilon}(\boldsymbol{x})))}$$

$$\alpha_2^{f_T\to f_S} = \|\mathbb{E}_{\boldsymbol{x}\sim\mathscr{D}}[\widehat{\Delta_{f_T\to f_T}(\boldsymbol{x})}\widehat{\Delta_{f_T\to f_S}(\boldsymbol{x})}^\top]\|_F$$