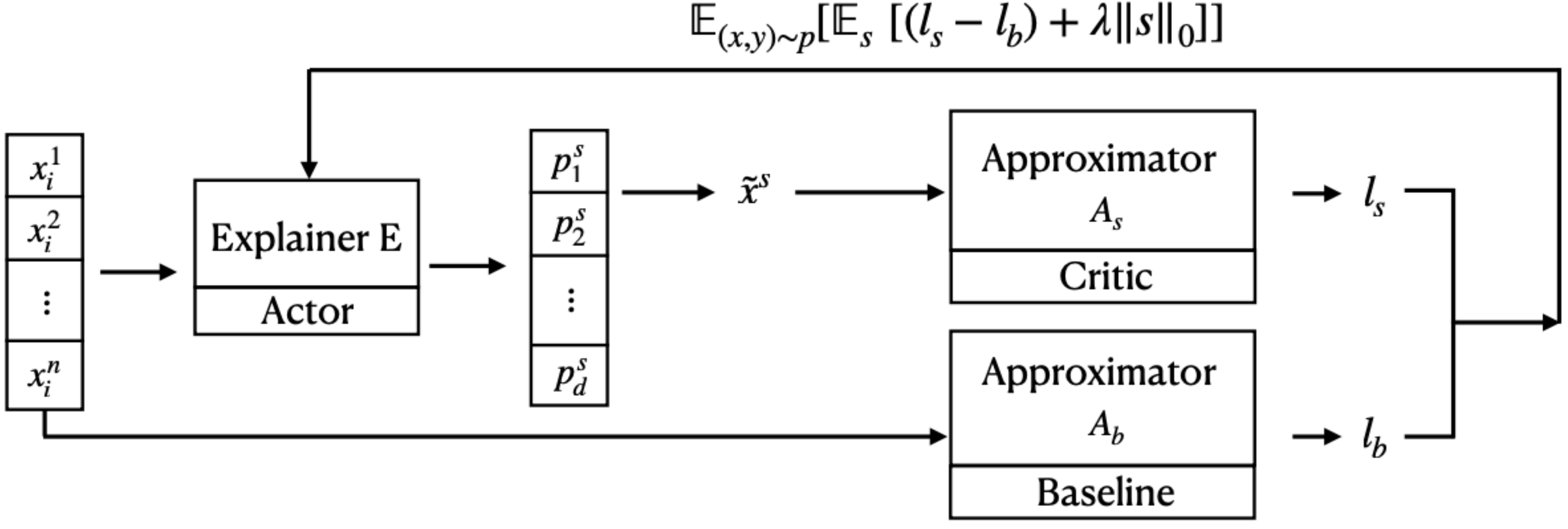
Related Work

INVASE (Actor-Critic Based)

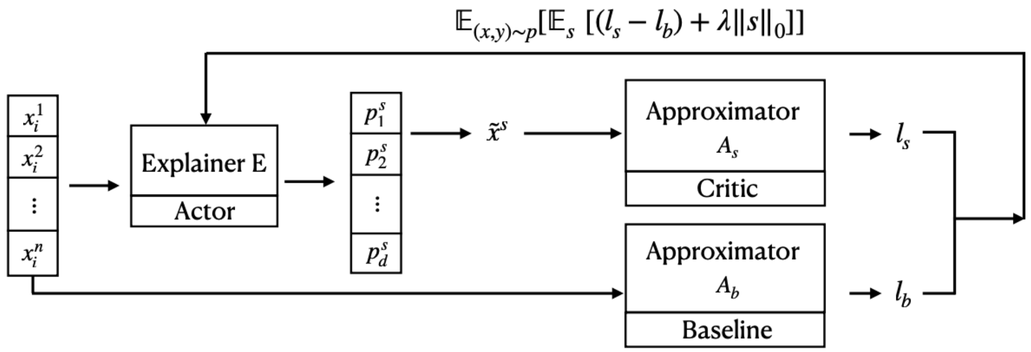

$\mathbb{E}_{(x,y)\sim p}[\mathbb{E}_s\ [(l_s - l_b) + \lambda\|s\|_0]]$
x_i^1
x_i^2
x_i^n
Explainer E
Actor
p_1^s
p_2^s
p_d^s
$\tilde{x}^s$
Approximator
A_s
Critic
Approximator
A_b
Baseline
l_s
l_b

- Actor (Explainer): Neural netwrok Model
- Critic: Neural netwrok Model that use partial feature values to approximate original black-box model

✓ Achieve nice performance on synthetic datasets in terms of TPR & FDR

Related Work

INVASE (Actor-Critic Based)



Related Work

INVASE (Actor-Critic Based)