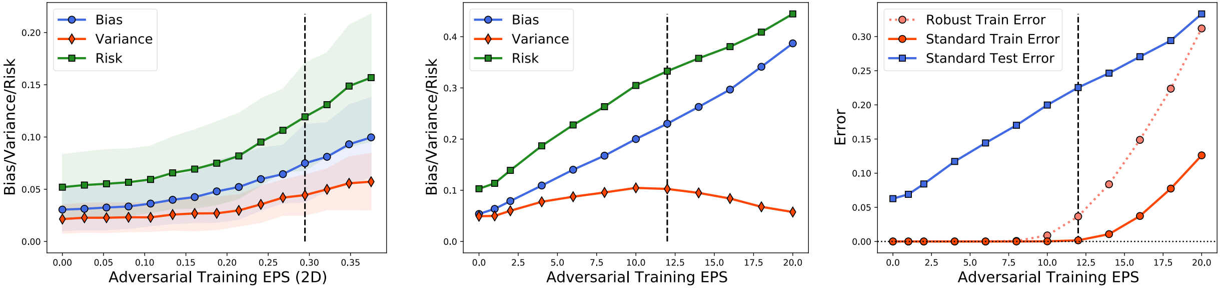

(a) Bias/variance/risk for *2D box example*.

(b) Bias/variance/risk and training/test error for *CIFAR10*.

Figure 1. Measuring performance for ℓ_∞-adversarial training (with increasing perturbation size) on the *2D box* dataset (1(a)) and *CIFAR10* dataset (1(b)). **Standard error** means the error rate on clean samples, and **robust error** means the error rate on adversarially perturbed samples. The *vertical dashed line* corresponds to the robust training error of the adversarially trained model is larger than 2%. (**a**) Evaluating the bias, variance, and risk for the ℓ_∞-adversarially trained model (fully connected network) on *2D box* dataset. (**b**) Evaluating bias, variance, risk and robust training error, standard training/test error for the ℓ_∞-adversarially trained model (WideResNet-28-10) on *CIFAR10* dataset.