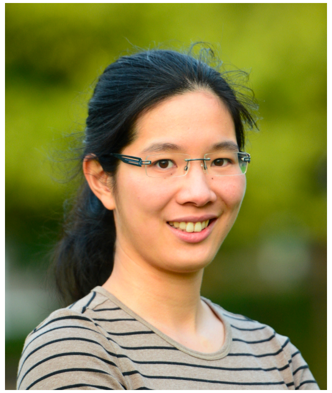
Understanding and Mitigating the Tradeoff Between Robustness and Accuracy



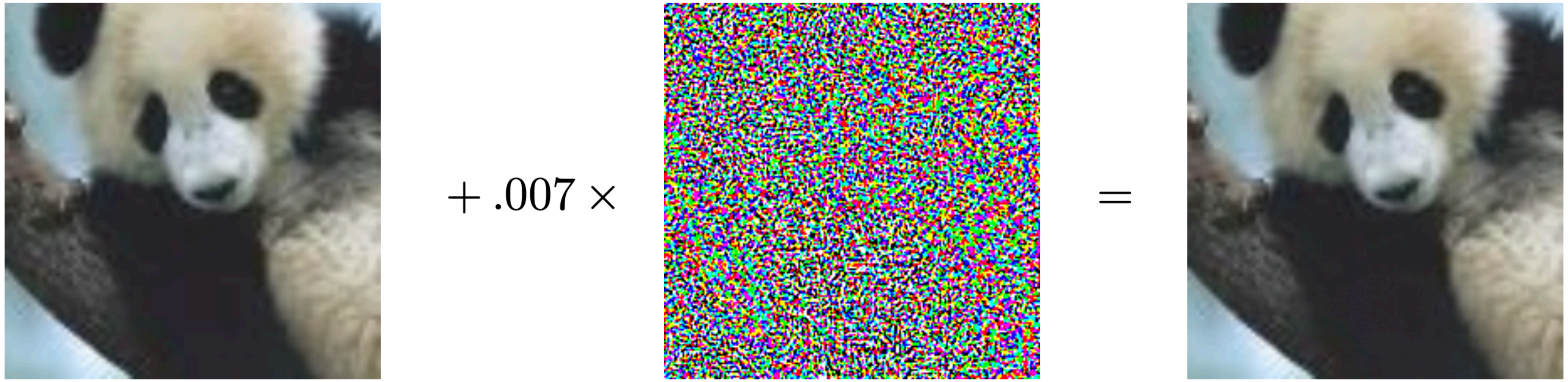


Aditi Raghunathan* **Sang Michael Xie*** Fanny Yang John C. Duchi Percy Liang

Stanford University

Adversarial examples

- Standard training leads to models that are not robust



"panda"
57.7% confidence

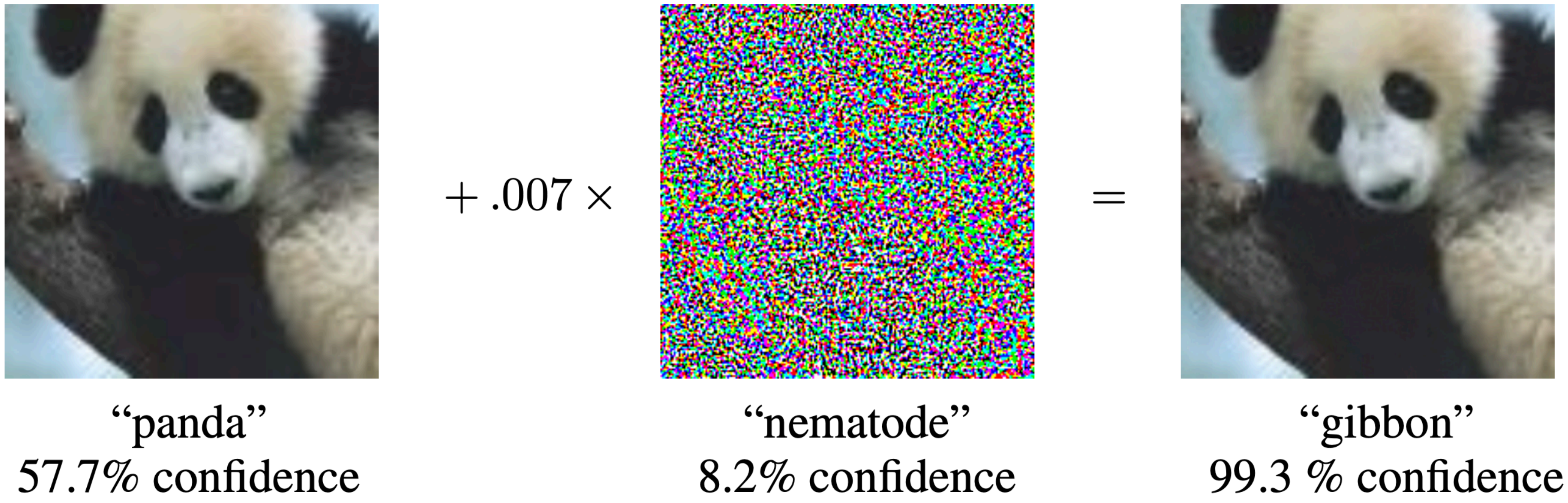
"nematode"
8.2% confidence

"gibbon"
99.3 % confidence

[Goodfellow et al. 2015]

Adversarial examples

- Standard training leads to models that are not robust



[Goodfellow et al. 2015]

- **Adversarial training** is a popular approach to improve robustness
- It **augments** the training set on-the-fly with adversarial examples

Adversarial training increases standard error

CIFAR-10

Method	Robust Accuracy
Standard Training	0%
TRADES Adversarial Training (Zhang et al. 2019)	**55.4%**

Robust Accuracy**:** % of test examples misclassified after an ℓ_∞-bounded adversarial perturbation

Adversarial training increases standard error

CIFAR-10

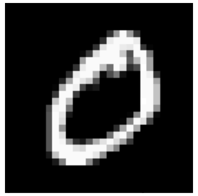
Method	Robust Accuracy	Standard Accuracy
Standard Training	0%	**95.2%**
TRADES Adversarial Training (Zhang et al. 2019)	**55.4%**	**84.0%**

Robust Accuracy: % of test examples misclassified after an ℓ_∞-bounded adversarial perturbation

Why is there a **tradeoff** between robustness and accuracy? We only augmented with more data!

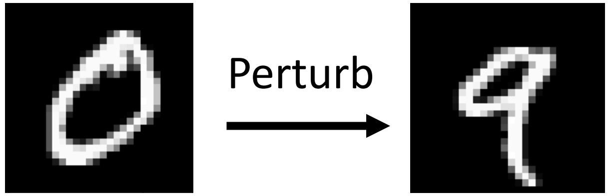
Prior hypotheses for the tradeoff

- Optimal predictor not robust to adversarial perturbations [Tsipras et al. 2019]



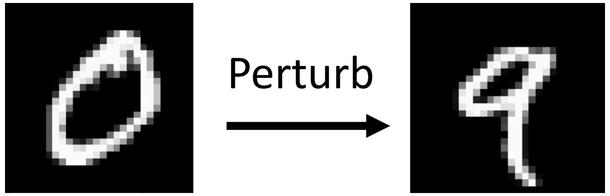
Prior hypotheses for the tradeoff

- Optimal predictor not robust to adversarial perturbations [Tsipras et al. 2019]

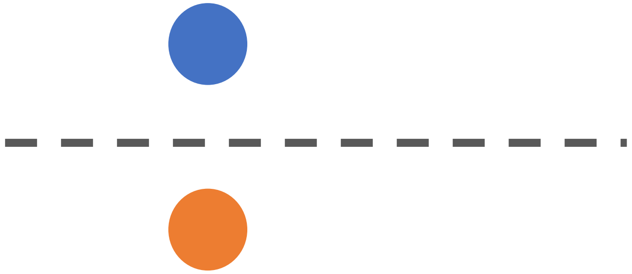


Prior hypotheses for the tradeoff

- Optimal predictor not robust to adversarial perturbations [Tsipras et al. 2019]

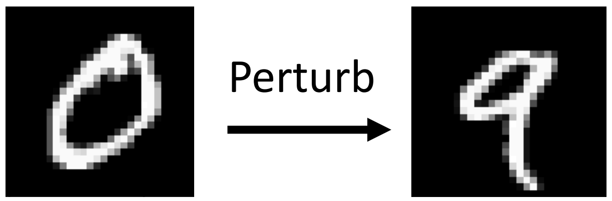


- Hypothesis class not expressive enough [Nakkiran et al. 2019]

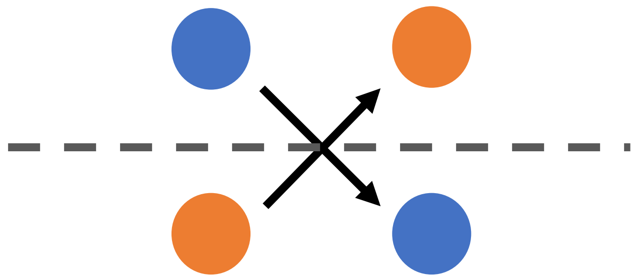


Prior hypotheses for the tradeoff

- Optimal predictor not robust to adversarial perturbations [Tsipras et al. 2019]

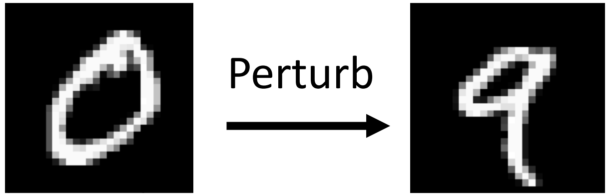


- Hypothesis class not expressive enough [Nakkiran et al. 2019]

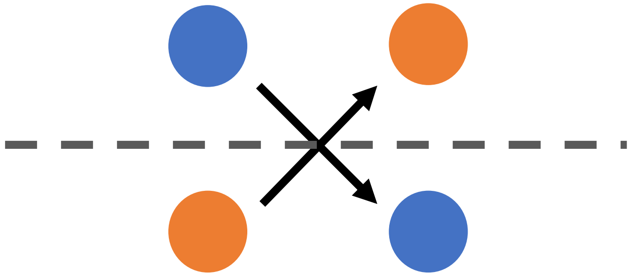


Prior hypotheses for the tradeoff

- Optimal predictor not robust to adversarial perturbations [Tsipras et al. 2019]



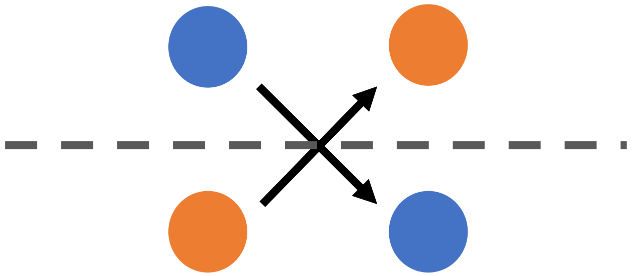
- Hypothesis class not expressive enough [Nakkiran et al. 2019]



These hypotheses suggest a tradeoff even in the infinite data limit…

Prior hypotheses for the tradeoff

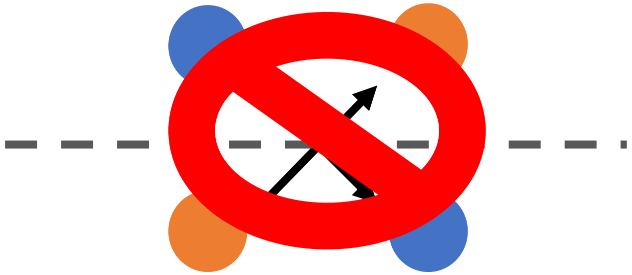
- Optimal predictor not robust to adversarial perturbations [Tsipras et al. 2019]
  - But typical perturbations are imperceptible, robustness should be possible
- Hypothesis class not expressive enough [Nakkiran et al. 2019]



These hypotheses suggest a tradeoff even in the infinite data limit…

Prior hypotheses for the tradeoff

- Optimal predictor not robust to adversarial perturbations [Tsipras et al. 2019]
  - But typical perturbations are imperceptible, robustness should be possible
- Hypothesis class not expressive enough [Nakkiran et al. 2019]
  - But neural networks highly expressive, reaches 100% std and robust training accuracy



These hypotheses suggest a tradeoff even in the infinite data limit…

Prior hypotheses for the tradeoff

More realistic settings:

- Optimal predictor not robust to adversarial perturbations [Tsipras et al. 2019]
  - But typical perturbations are imperceptible, robustness should be possible
- Hypothesis class not expressive enough [Nakkiran et al. 2019]
  - But neural networks highly expressive, reaches 100% std and robust training accuracy

Consistent

These hypotheses suggest a tradeoff even in the infinite data limit…

Prior hypotheses for the tradeoff

More realistic settings:

- Optimal predictor not robust to adversarial perturbations [Tsipras et al. 2019]
  - But typical perturbations are imperceptible, robustness should be possible

Consistent

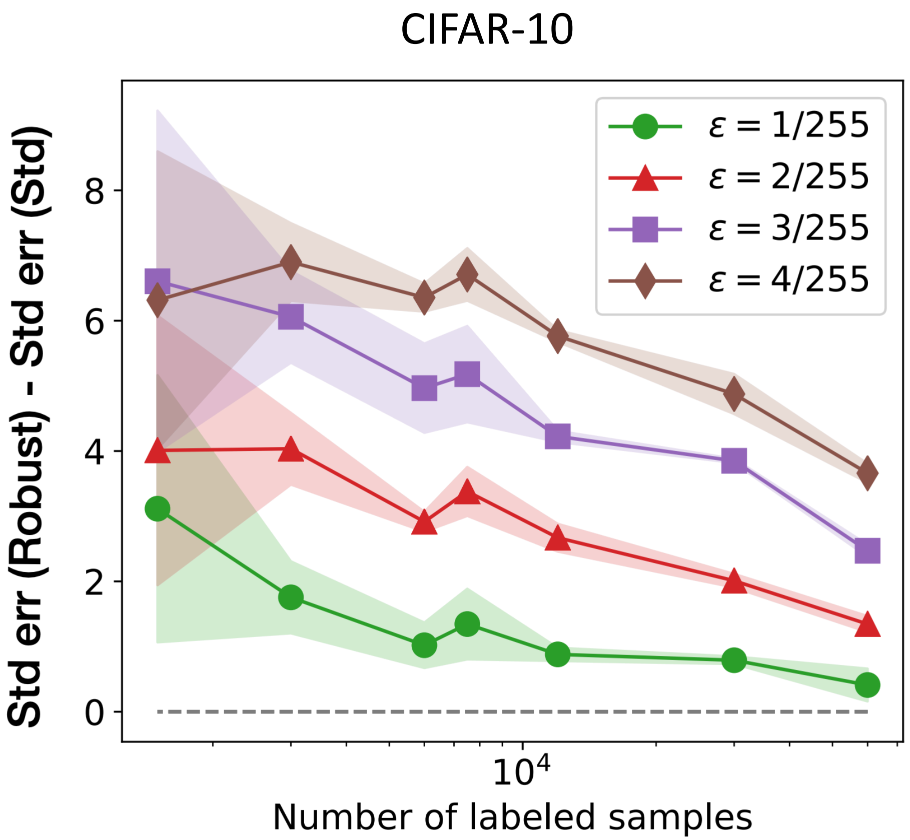
- Hypothesis class not expressive enough [Nakkiran et al. 2019]
  - But neural networks highly expressive, reaches 100% std and robust training accuracy

Well-specified

These hypotheses suggest a tradeoff even in the infinite data limit…

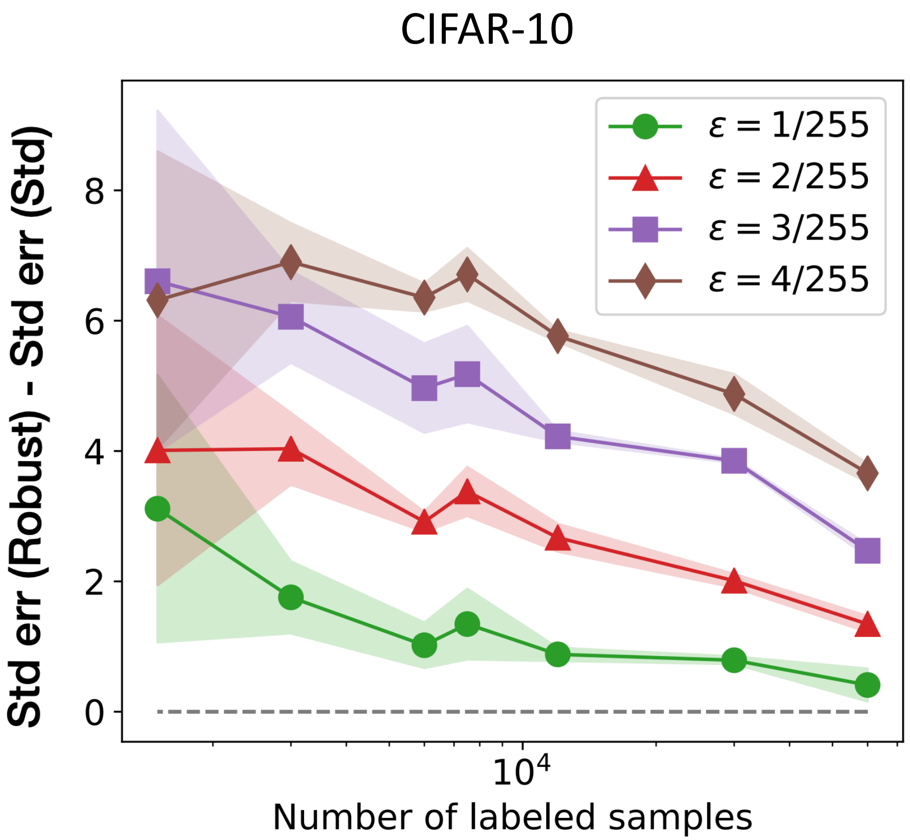
No tradeoff with infinite data

- Observations
  - Gap between robust and standard accuracies are large for small data regime
  - Gap decreases with labeled sample size



No tradeoff with infinite data

- Observations
  - Gap between robust and standard accuracies are large for small data regime
  - Gap decreases with labeled sample size
- We ask: if we have consistent perturbations + well-specified model family (no inherent tradeoff), why do we observe a tradeoff in practice?



Results overview

- Characterize how **training with consistent extra data can increase standard error** even in well-specified noiseless linear regression
  - Analysis suggests robust self-training to mitigate tradeoff [Carmon 2019, Najafi 2019, Uesato 2019]

Results overview

- Characterize how **training with consistent extra data can increase standard error** even in well-specified noiseless linear regression
  - Analysis suggests robust self-training to mitigate tradeoff [Carmon 2019, Najafi 2019, Uesato 2019]

- Prove that robust self-training **(RST) improves robust error without hurting standard error** in linear setting with unlabeled data

Results overview

- Characterize how **training with consistent extra data can increase standard error** even in well-specified noiseless linear regression
  - Analysis suggests robust self-training to mitigate tradeoff [Carmon 2019, Najafi 2019, Uesato 2019]
- Prove that robust self-training **(RST) improves robust error without hurting standard error** in linear setting with unlabeled data
- Empirically, RST improves robust **and** standard error across different adversarial training algorithms and adversarial perturbation types

Noiseless linear regression

- Model: $y = x^\top \theta^*$ Well-specified

Noiseless linear regression

- Model: $y = x^{\top}\theta^*$ Well-specified

- Standard data: $X_{std} \in \mathbb{R}^{n \times d}$, $y_{std} = X_{std}\theta^*$, $n \ll d$ (overparameterized)

Noiseless linear regression

- Model: $y = x^\top \theta^*$ Well-specified
- Standard data: $X_{std} \in \mathbb{R}^{n \times d}$, $y_{std} = X_{std}\theta^*$, $n \ll d$ (overparameterized)
- Extra data (adv examples): $X_{ext} \in \mathbb{R}^{m \times d}$, $y_{ext} = X_{ext}\theta^*$ Consistent

Noiseless linear regression

- Model: $y = x^\top \theta^*$ Well-specified
- Standard data: $X_{std} \in \mathbb{R}^{n \times d}$, $y_{std} = X_{std}\theta^*$, $n \ll d$ (overparameterized)
- Extra data (adv examples): $X_{ext} \in \mathbb{R}^{m \times d}$, $y_{ext} = X_{ext}\theta^*$ Consistent
- We study min-norm interpolants
  - $\theta_{std} = \operatorname{argmin}_\theta \{\|\theta\|_2 : X_{std}\theta = y_{std}\}$
  - $\theta_{aug} = \operatorname{argmin}_\theta \{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$

Noiseless linear regression

- Model: $y = x^{\top}\theta^{*}$ Well-specified
- Standard data: $X_{std} \in \mathbb{R}^{n \times d}$, $y_{std} = X_{std}\theta^{*}$, $n \ll d$ (overparameterized)
- Extra data (adv examples): $X_{ext} \in \mathbb{R}^{m \times d}$, $y_{ext} = X_{ext}\theta^{*}$ Consistent
- We study min-norm interpolants
  - $\theta_{std} = \text{argmin}_{\theta}\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$
  - $\theta_{aug} = \text{argmin}_{\theta}\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$
- Standard error: $(\theta - \theta^{*})^{\top}\Sigma(\theta - \theta^{*})$ for population covariance Σ

Example: when extra data hurts standard error

$\theta_{std} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$

$\theta_{aug} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$

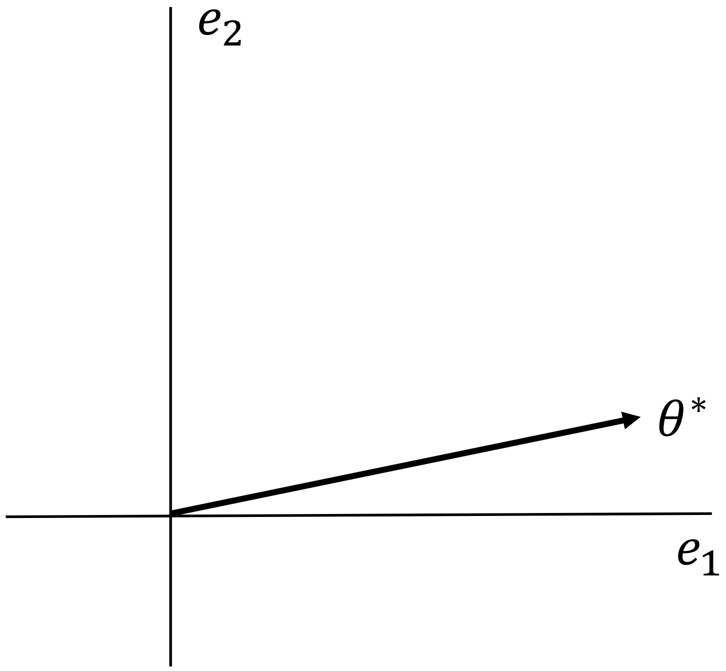
- Min-norm interpolants + noiseless: recover θ^* exactly in span of training data

Example: when extra data hurts standard error

$\theta_{std} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$

$\theta_{aug} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$

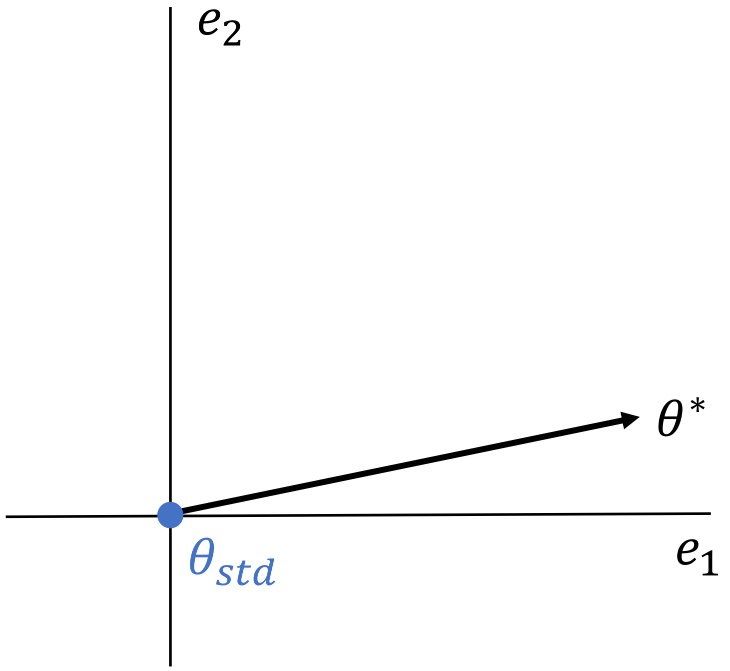
- Min-norm interpolants + noiseless: recover θ^* exactly in span of training data
- Suppose null space of X_{std} is $[e_1, e_2]$



Example: when extra data hurts standard error

$\theta_{std} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$

$\theta_{aug} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$

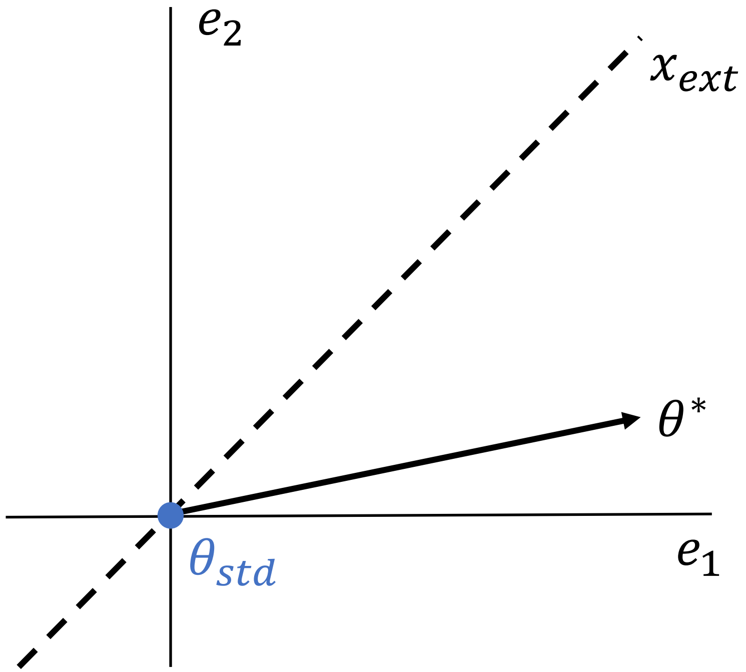


- Min-norm interpolants + noiseless: recover θ^* exactly in span of training data
- Suppose null space of X_{std} is $[e_1, e_2]$

Example: when extra data hurts standard error

$\theta_{std} = \text{argmin}_{\theta}\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$

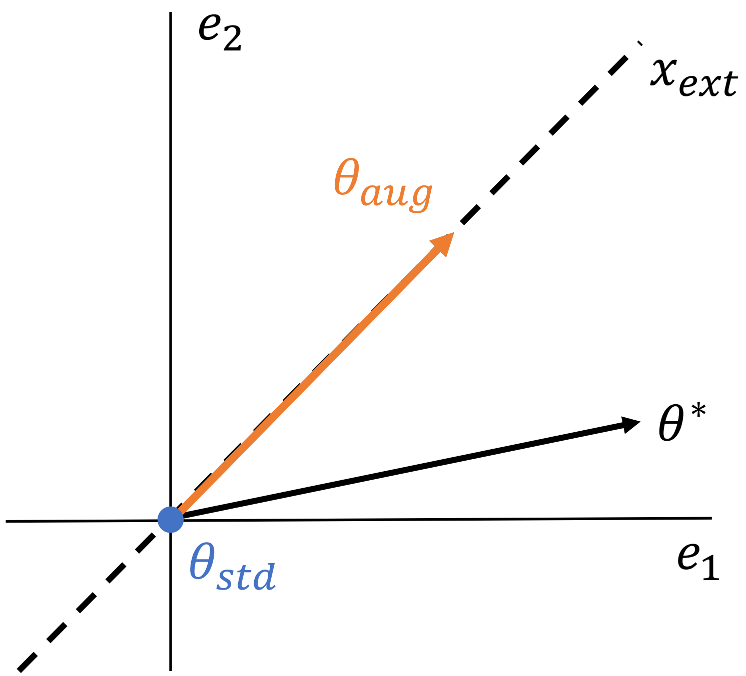
$\theta_{aug} = \text{argmin}_{\theta}\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$



- Min-norm interpolants + noiseless: recover θ^* exactly in span of training data
- Suppose null space of X_{std} is $[e_1, e_2]$

Example: when extra data hurts standard error

$$\theta_{std} = \operatorname{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$$
$$\theta_{aug} = \operatorname{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$$

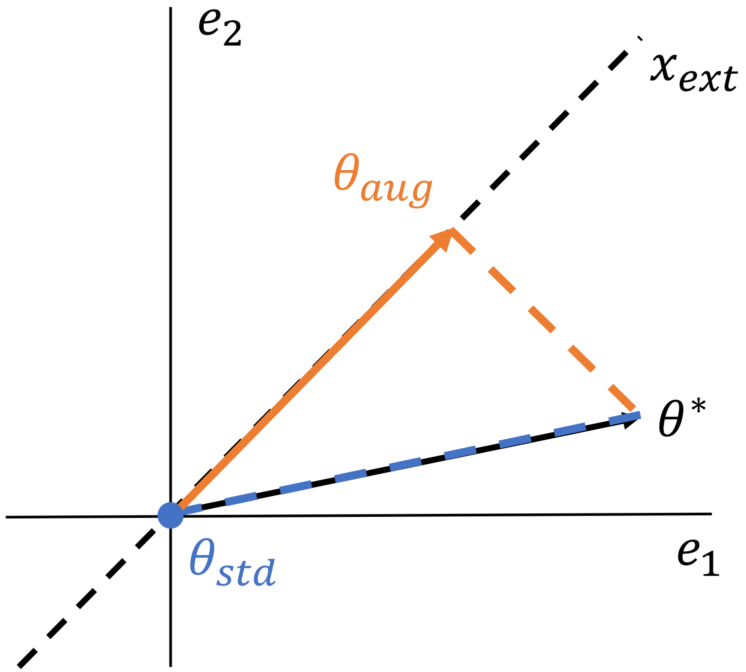


- Min-norm interpolants + noiseless: recover θ^* exactly in span of training data
- Suppose null space of X_{std} is $[e_1, e_2]$
- θ_{aug} fits θ^* in x_{ext} direction, 0 otherwise

Example: when extra data hurts standard error

$\theta_{std} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$

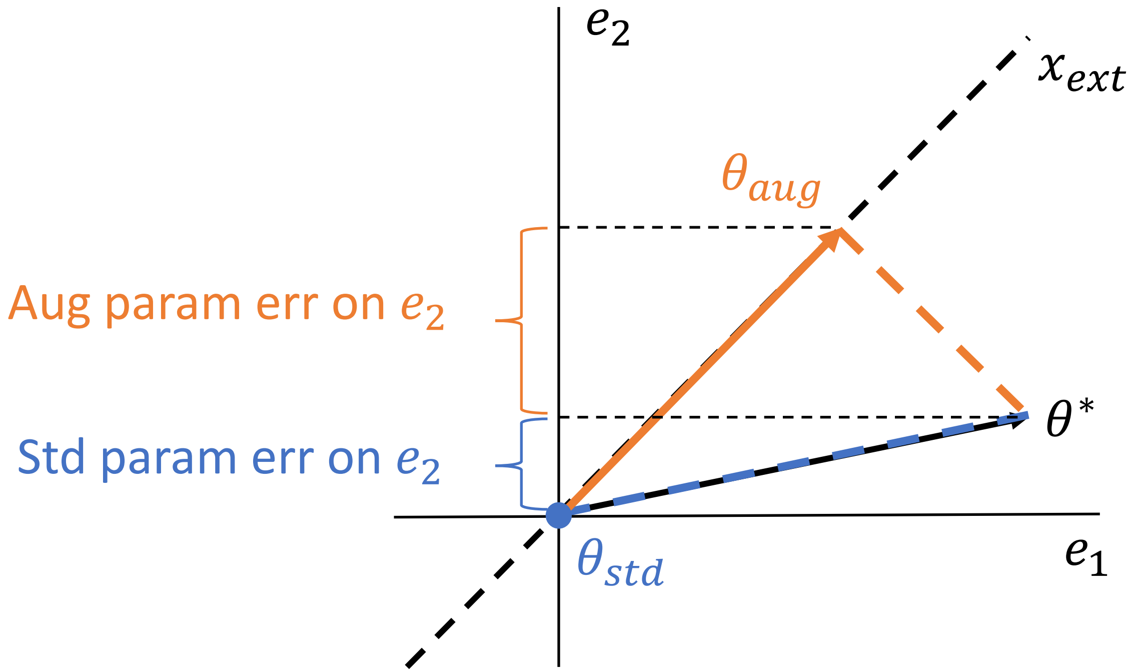
$\theta_{aug} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$



- Min-norm interpolants + noiseless: recover θ^* exactly in span of training data
- Suppose null space of X_{std} is $[e_1, e_2]$
- θ_{aug} fits θ^* in x_{ext} direction, 0 otherwise

Example: when extra data hurts standard error

$\theta_{std} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$
$\theta_{aug} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$

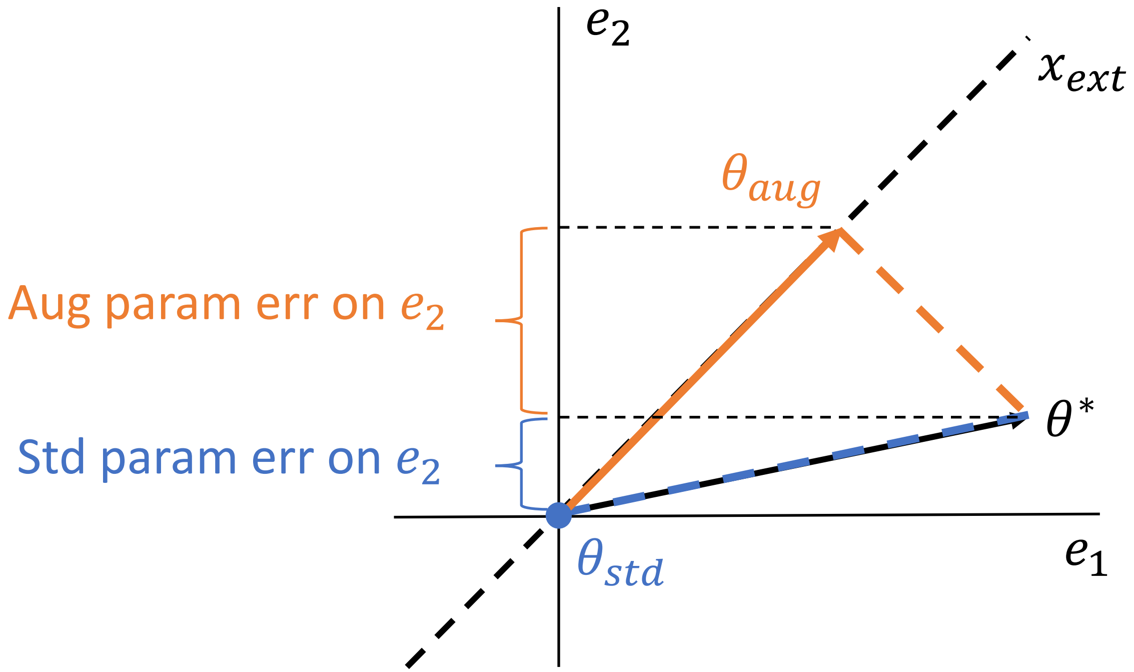


- Min-norm interpolants + noiseless: recover θ^* exactly in span of training data
- Suppose null space of X_{std} is $[e_1, e_2]$
- θ_{aug} fits θ^* in x_{ext} direction, 0 otherwise
- If Σ has high weight on e_2 direction, errors in e_2 are more costly $\Rightarrow$ augmented estimator has higher error

Example: when extra data hurts standard error

$\theta_{std} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$

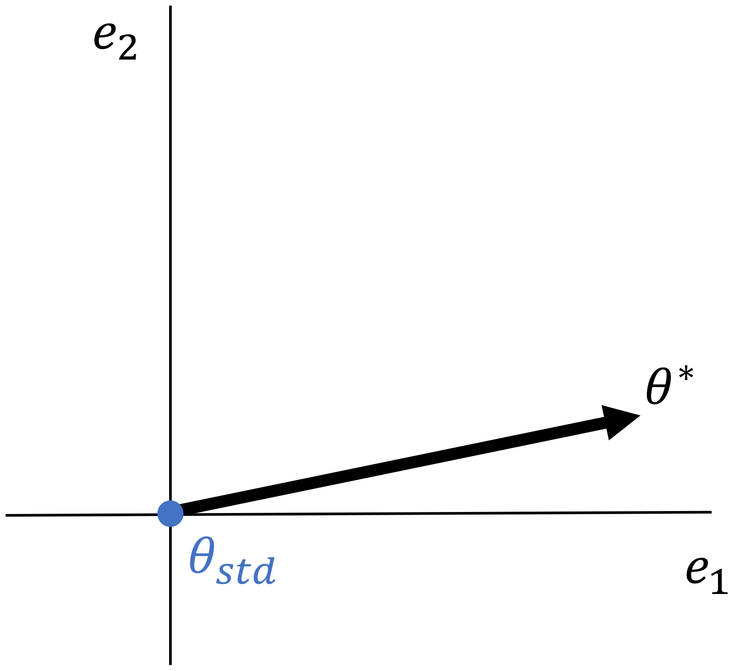
$\theta_{aug} = \text{argmin}_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{ext}\theta = y_{ext}\}$



- Min-norm interpolants + noiseless: recover θ^* exactly in span of training data
- Suppose null space of X_{std} is $[e_1, e_2]$
- θ_{aug} fits θ^* in x_{ext} direction, 0 otherwise
- If Σ has high weight on e_2 direction, errors in e_2 are more costly $\Rightarrow$ augmented estimator has higher error
- The paper has exact characterization for noiseless linear regression setting

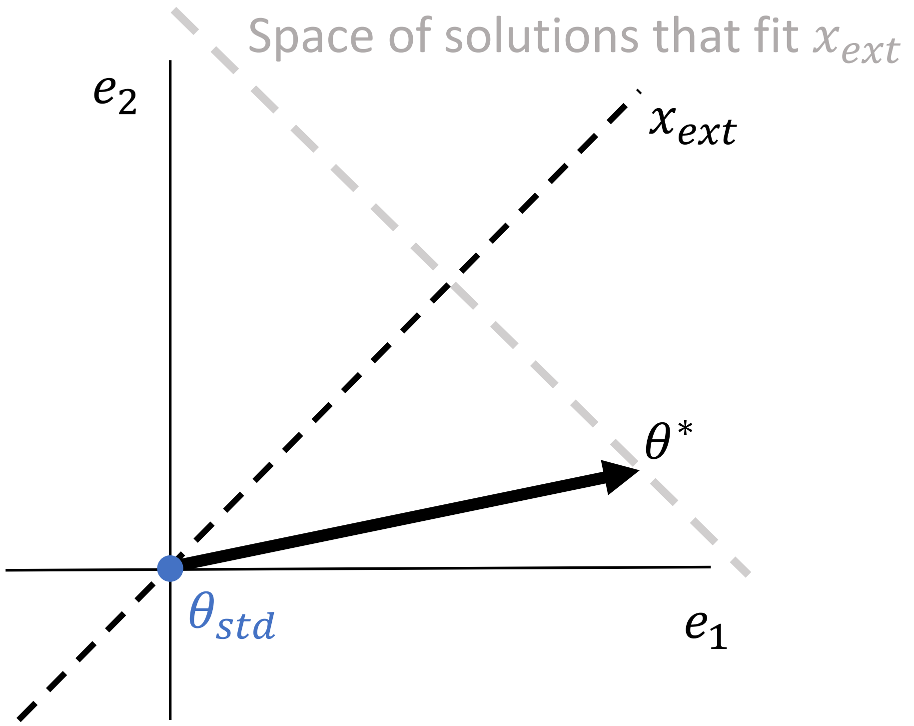
Mitigating the increase in error

- Suppose we know the population covariance Σ has high weight on e_2



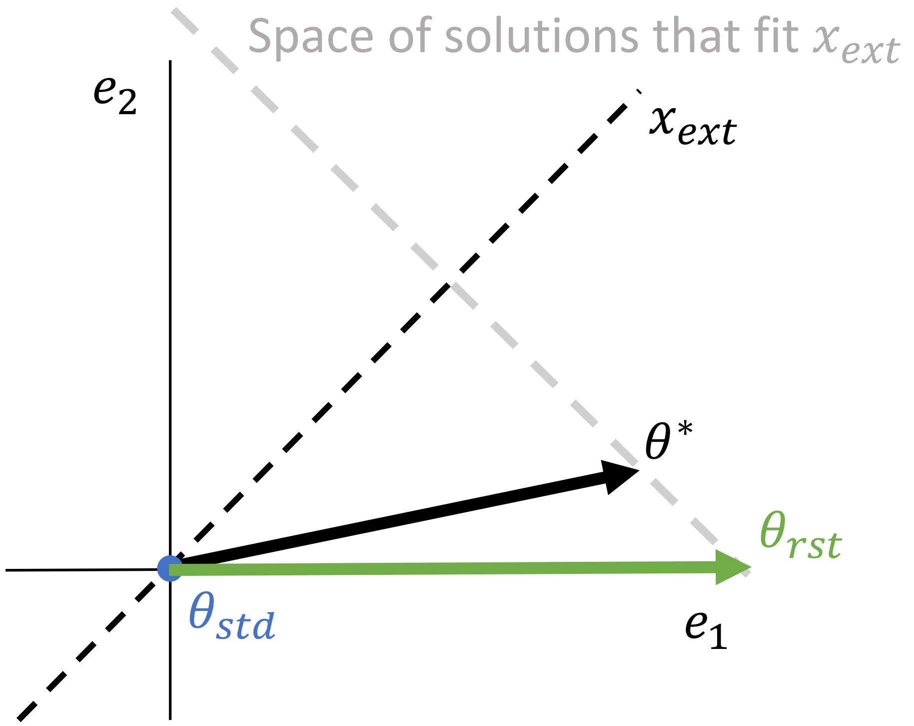
Mitigating the increase in error

- Suppose we know the population covariance Σ has high weight on e_2



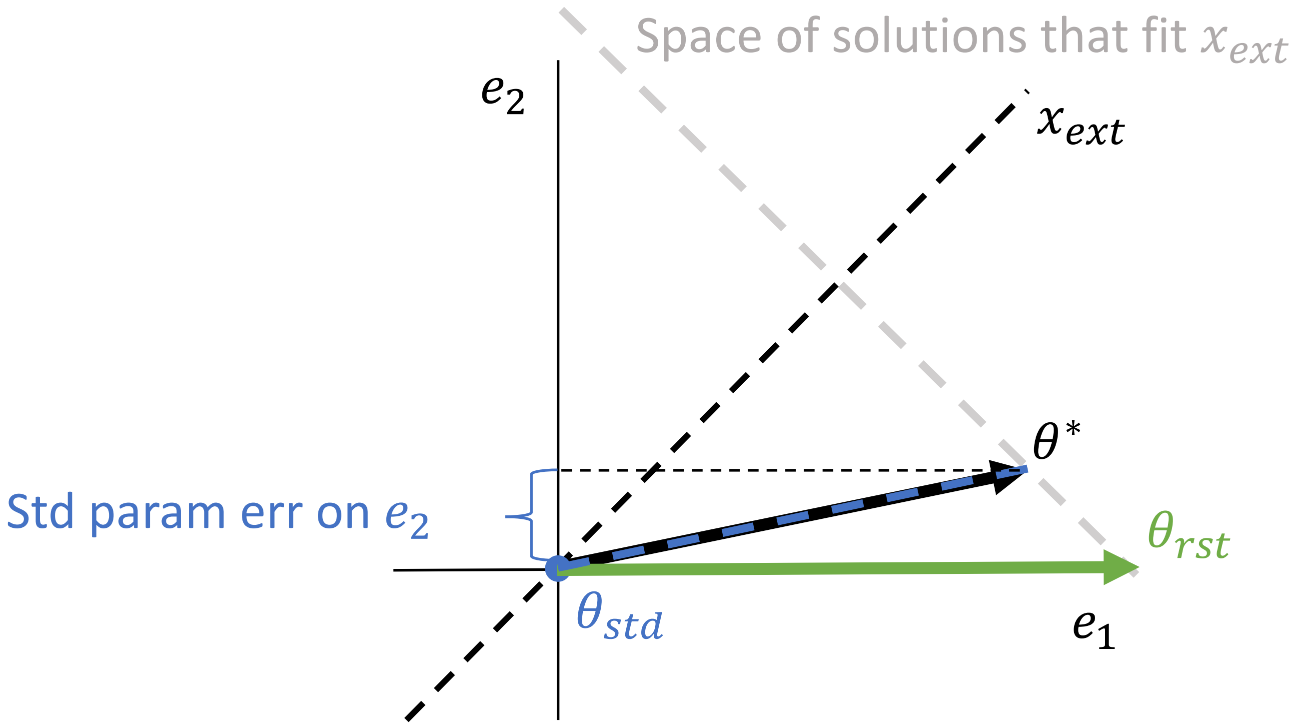
Mitigating the increase in error

- Suppose we know the population covariance Σ has high weight on e_2
- To mitigate error, regularize toward θ_{std} on e_2 component



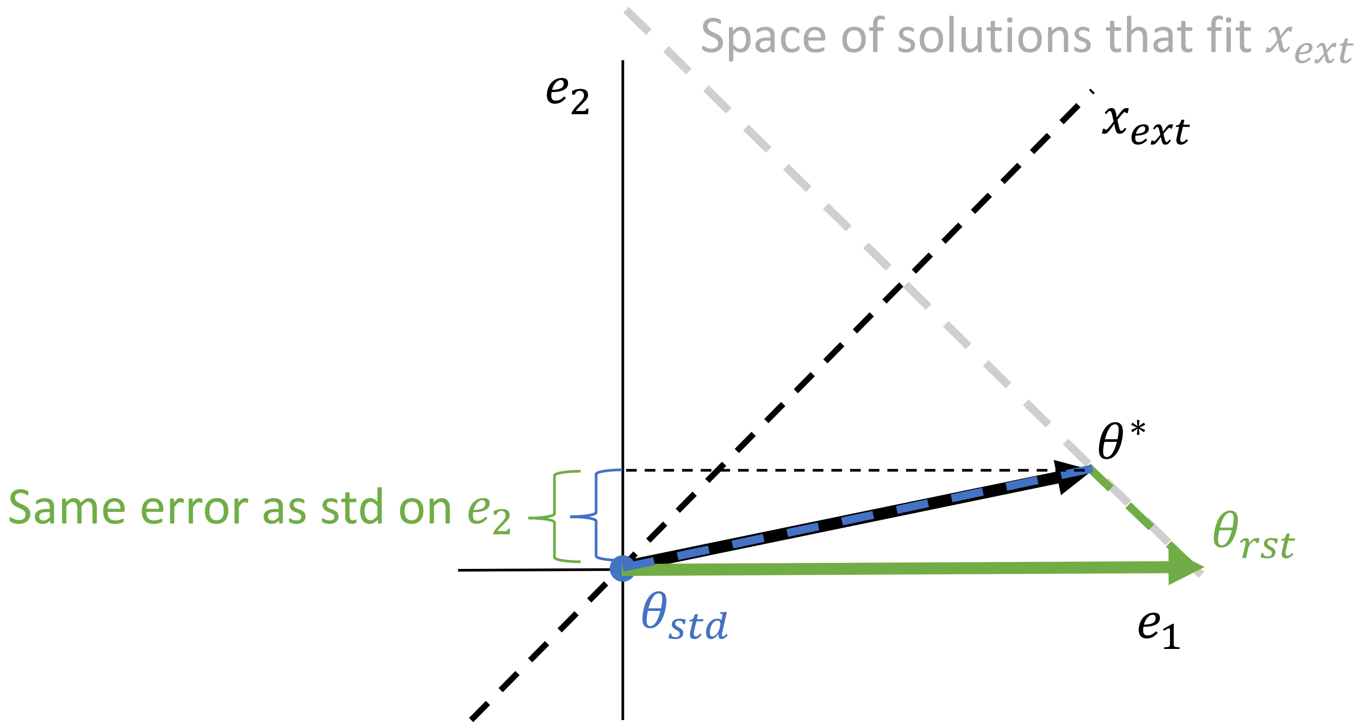
Mitigating the increase in error

- Suppose we know the population covariance Σ has high weight on e_2
- To mitigate error, regularize toward θ_{std} on e_2 component



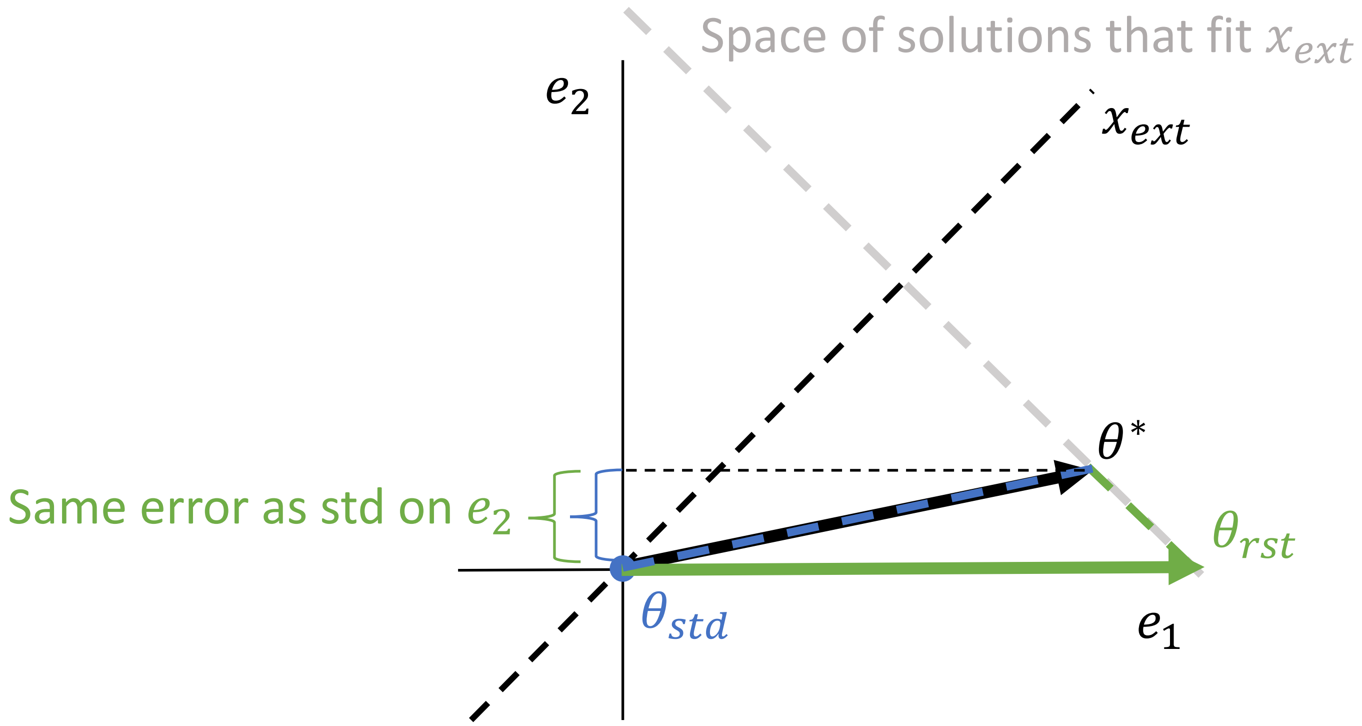
Mitigating the increase in error

- Suppose we know the population covariance Σ has high weight on e_2
- To mitigate error, regularize toward θ_{std} on e_2 component



Mitigating the increase in error

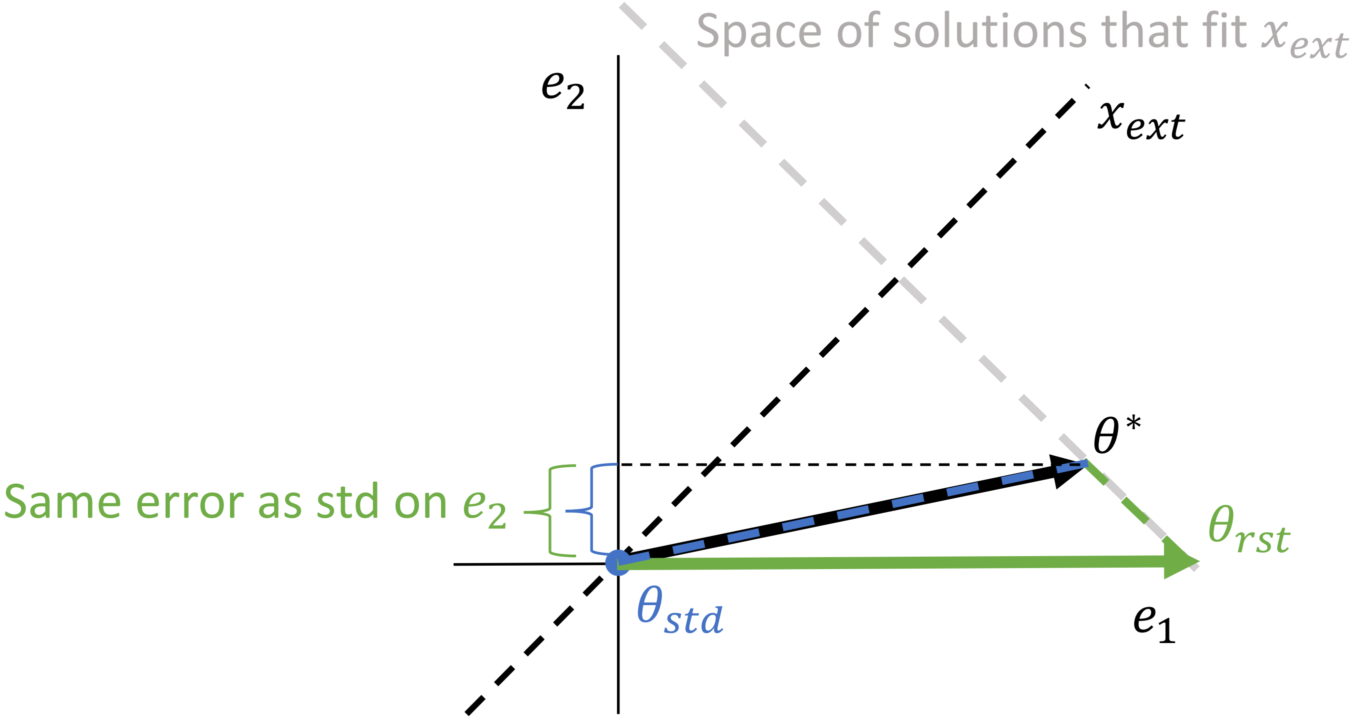
- Suppose we know the population covariance Σ has high weight on e_2
- To mitigate error, regularize toward θ_{std} on e_2 component



- Idea: Use **unlabeled data** to estimate Σ

Mitigating the increase in error

- Suppose we know the population covariance Σ has high weight on e_2
- To mitigate error, regularize toward θ_{std} on e_2 component



- Idea: Use **unlabeled data** to estimate Σ

We show this is exactly Robust Self-Training!

Robust Self-Training (RST)

- Recent semi-supervised algorithm that can be applied on top of existing adversarial training methods (Carmon et al., Najafi et al., Uesato et al.)
- Labeled examples (x, y)

Components of RST

	Standard	**Robust (extra data $x_{ext} = x_{adv}$)**
Labeled	Fit (x, y)	Fit (x_{adv}, y)

Robust Self-Training (RST)

- Recent semi-supervised algorithm that can be applied on top of existing adversarial training methods (Carmon et al., Najafi et al., Uesato et al.)
- Labeled examples (x, y)
- Unlabeled examples $\tilde{x}$ $\rightarrow$ standard predictor $\rightarrow$ pseudo-labels $\tilde{y}$

Components of RST

	Standard	Robust (extra data $x_{ext} = x_{adv}$)
Labeled	Fit (x, y)	Fit (x_{adv}, y)
Unlabeled	Fit $(\tilde{x}, \tilde{y})$	Fit $(\tilde{x}_{adv}, \tilde{y})$

Robust Self-Training (RST)

- Recent semi-supervised algorithm that can be applied on top of existing adversarial training methods (Carmon et al., Najafi et al., Uesato et al.)
- Labeled examples (x, y)
- Unlabeled examples $\tilde{x} \rightarrow$ standard predictor $\rightarrow$ pseudo-labels $\tilde{y}$

Components of RST

	Standard	Robust (extra data $x_{ext} = x_{adv}$)
Labeled	Fit (x, y)	Fit (x_{adv}, y)
Unlabeled	Fit $(\tilde{x}, \tilde{y})$	Fit $(\tilde{x}_{adv}, \tilde{y})$

Theorem (informal): for noiseless linear regression, **RST always improves both standard and robust errors**

RST mitigates tradeoff in adversarial training

- RST mitigates tradeoff for adv. training with both TRADES and PG-AT

CIFAR-10

Method	Robust Accuracy	Standard Accuracy
Standard Training	0%	95.2%
PG-AT (Madry et al. 2018)	45.8%	87.3%
TRADES (Zhang et al. 2019)	55.4%	84.0%

RST mitigates tradeoff in adversarial training

- RST mitigates tradeoff for adv. training with both TRADES and PG-AT

CIFAR-10

Method	Robust Accuracy	Standard Accuracy
Standard Training	0%	95.2%
PG-AT (Madry et al. 2018)	45.8%	87.3%
TRADES (Zhang et al. 2019)	55.4%	84.0%
RST + PG-AT	**58.5%**	**91.8%**
RST + TRADES	**63.1%**	**89.7%**

RST mitigates tradeoff in adversarial training

- RST mitigates tradeoff for adv. training with both TRADES and PG-AT
- Other semi-supervised approaches do not improve **standard** accuracy

CIFAR-10

Method	Robust Accuracy	Standard Accuracy
Standard Training	0%	95.2%
PG-AT (Madry et al. 2018)	45.8%	87.3%
TRADES (Zhang et al. 2019)	55.4%	84.0%
RST + PG-AT	**58.5%**	**91.8%**
RST + TRADES	**63.1%**	**89.7%**
Robust Consistency Training (Carmon et al. 2019)	56.5%	83.2%

RST mitigates tradeoff across perturbation types

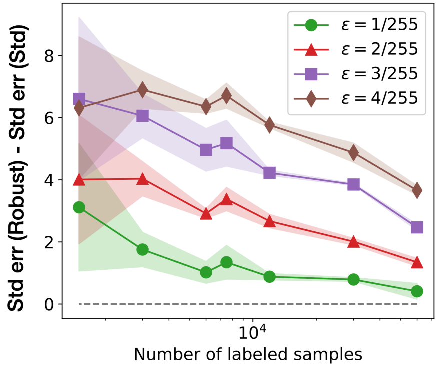
- Adversarial rotations + translations don't hurt standard error (Engstrom et al. 2019, Yang et al. 2019)
- Even in this case, RST improves both standard and robust error

CIFAR-10

Method	Robust Accuracy	Standard Accuracy
Standard Training	0.2%	94.6%
Worst-of-10	73.9%	95.0%
RST + Worst-of-10	**75.1%**	**95.8%**

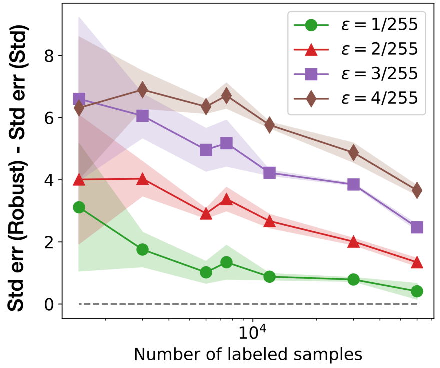
Takeaways

We characterize the tradeoff in noiseless linear regression in the more realistic setting of no inherent tradeoff.

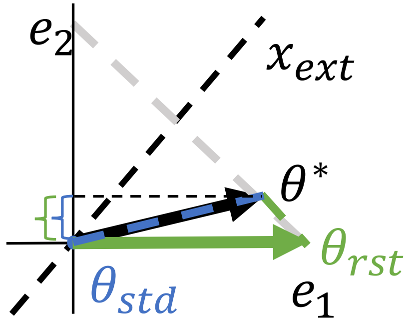


Takeaways

We characterize the tradeoff in noiseless linear regression in the more realistic setting of no inherent tradeoff.

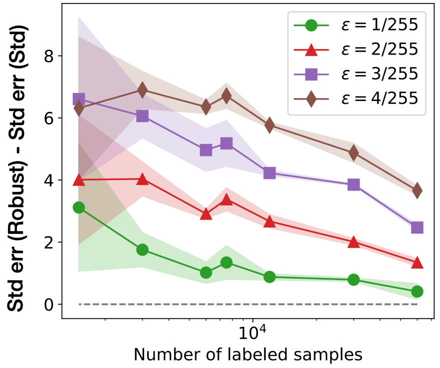


We show the effect of inductive bias in causing a tradeoff with finite data.

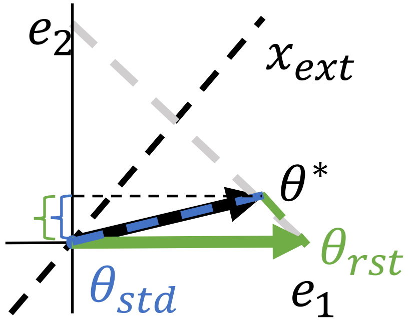


Takeaways

We characterize the tradeoff in noiseless linear regression in the more realistic setting of no inherent tradeoff.



We show the effect of inductive bias in causing a tradeoff with finite data.



Using unlabeled data, we can mitigate the tradeoff via robust self-training (RST).

Thanks!

This work was funded by an Open Philanthropy Project Award and NSF Frontier Award as part of the Center for Trustworthy Machine Learning (CTML).

AR was supported by Google Fellowship and Open Philanthropy AI Fellowship. FY was supported by the Institute for Theoretical Studies ETH Zurich and the Dr. Max Rossler and the Walter Haefner Foundation. FY and JCD were supported by the Office of Naval Research Young Investigator Awards. SMX was supported by an NDSEG Fellowship.

We thank the following people for valuable comments and discussions: Tengyu Ma, Yair Carmon, Ananya Kumar, Pang Wei Koh, Fereshte Khani, Shiori Sagawa, Karan Goel.