

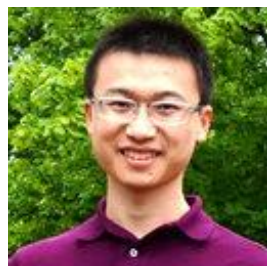


清华大学  
Tsinghua University

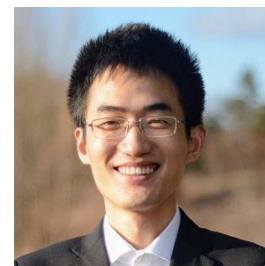
# Asymmetric Valleys: Beyond Sharp and Flat Local Minima



Haowei He<sup>1</sup>



Gao Huang<sup>2</sup>



Yang Yuan<sup>1</sup>

<sup>1</sup>Institute for Interdisciplinary Information Sciences, <sup>2</sup>Department of Automation

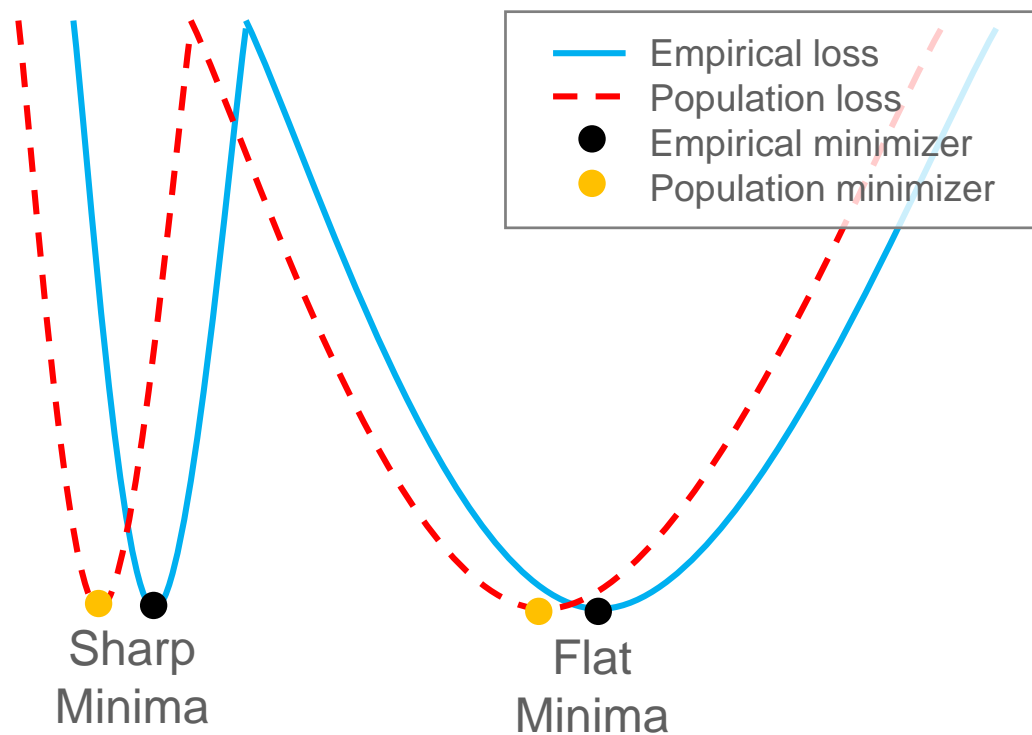
Tsinghua University

# Flat and Sharp minima



## Popular belief:

- Flat minima generalize better!\*



\* On large-batch training for deep learning: Generalization gap and sharp minima. ICLR, 2018.

# Flat and Sharp minima

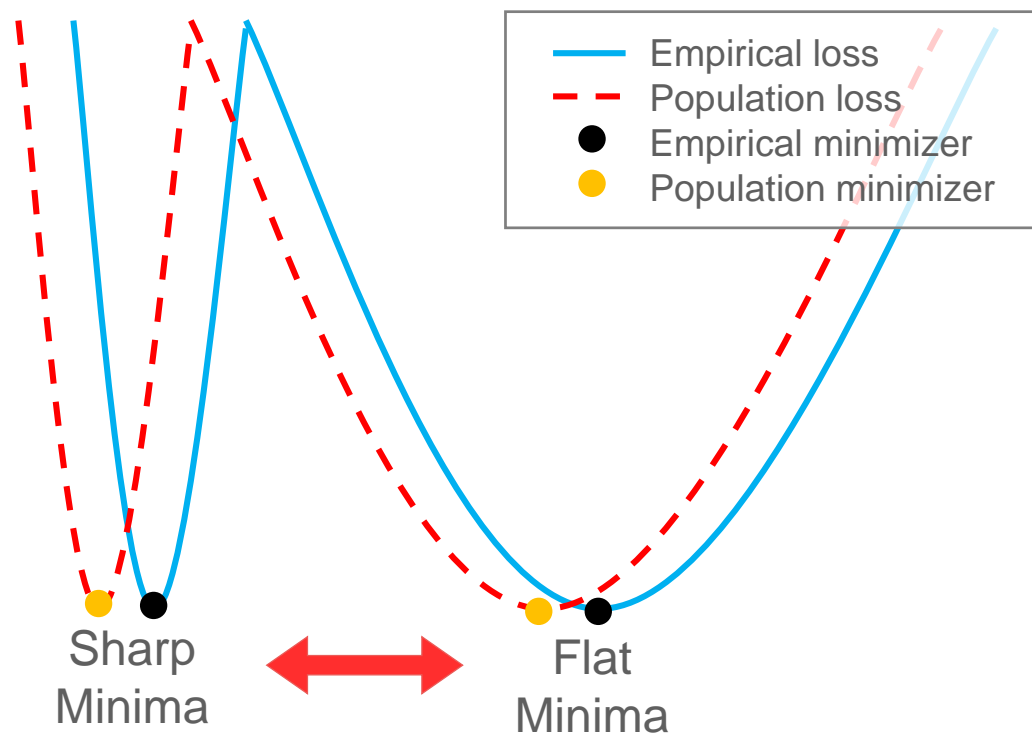


## Popular belief:

- Flat minima generalize better!

## Counter-examples:

- Flat and sharp minimum can convert to each other.\*



\* Sharp minima can generalize for deep nets. ICML, 2017.

# Flat and Sharp minima

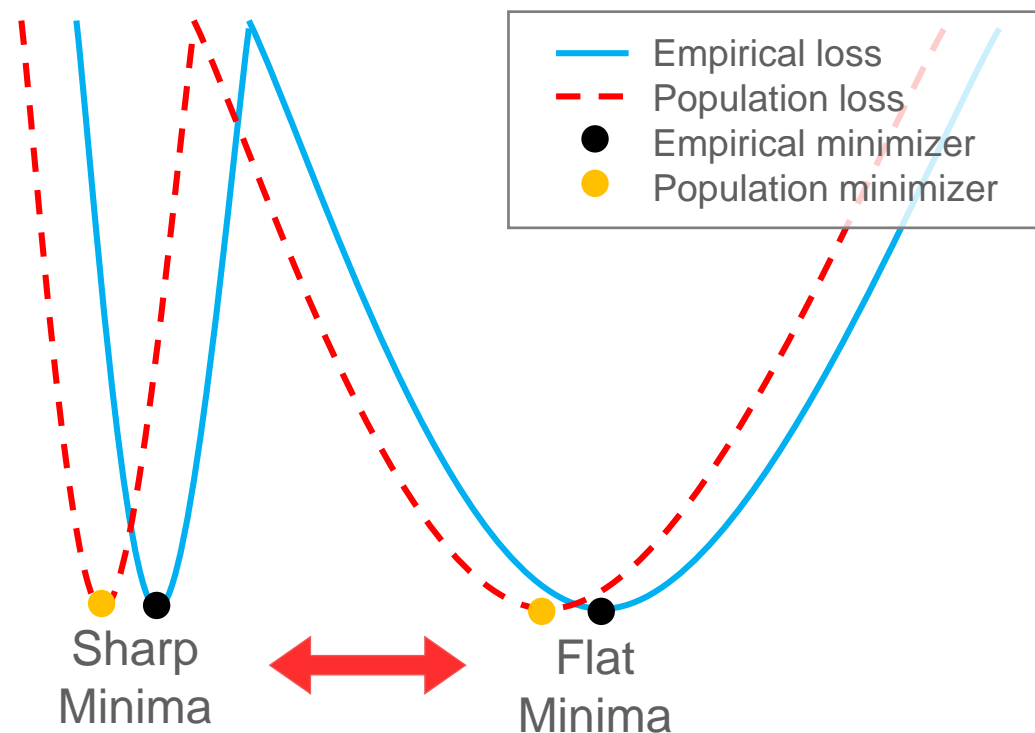


## Popular belief:

- Flat minima generalize better!

## Counter-examples:

- Flat and sharp minimum can convert to each other.\*
- Minima of modern deep networks are connected\*\*



\* Sharp minima can generalize for deep nets. ICML, 2017.

\*\* Essentially no barriers in neural network energy landscape. ICML, 2018.

# Flat and Sharp minima



清华大学  
Tsinghua University

Categorizing minima by flatness/sharpness might be an oversimplification!

# Flat and Sharp minima



清华大学  
Tsinghua University

Categorizing minima by flatness/sharpness might be an oversimplification!

In a minimum, the landscape might be **sharp** along some directions, but **flat** along other directions.

# Our Proposal: Asymmetric Valley

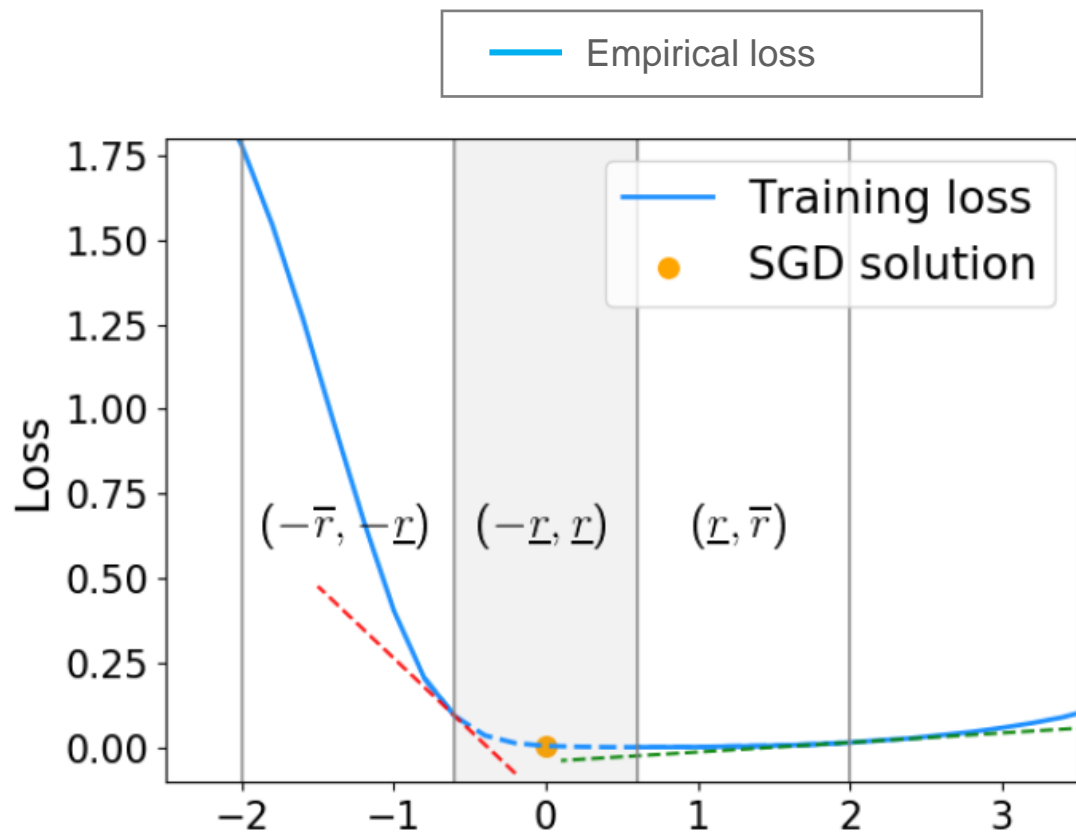


## Asymmetric Valley:

Loss grows fast on one side and slowly on the other side.

### Definition:

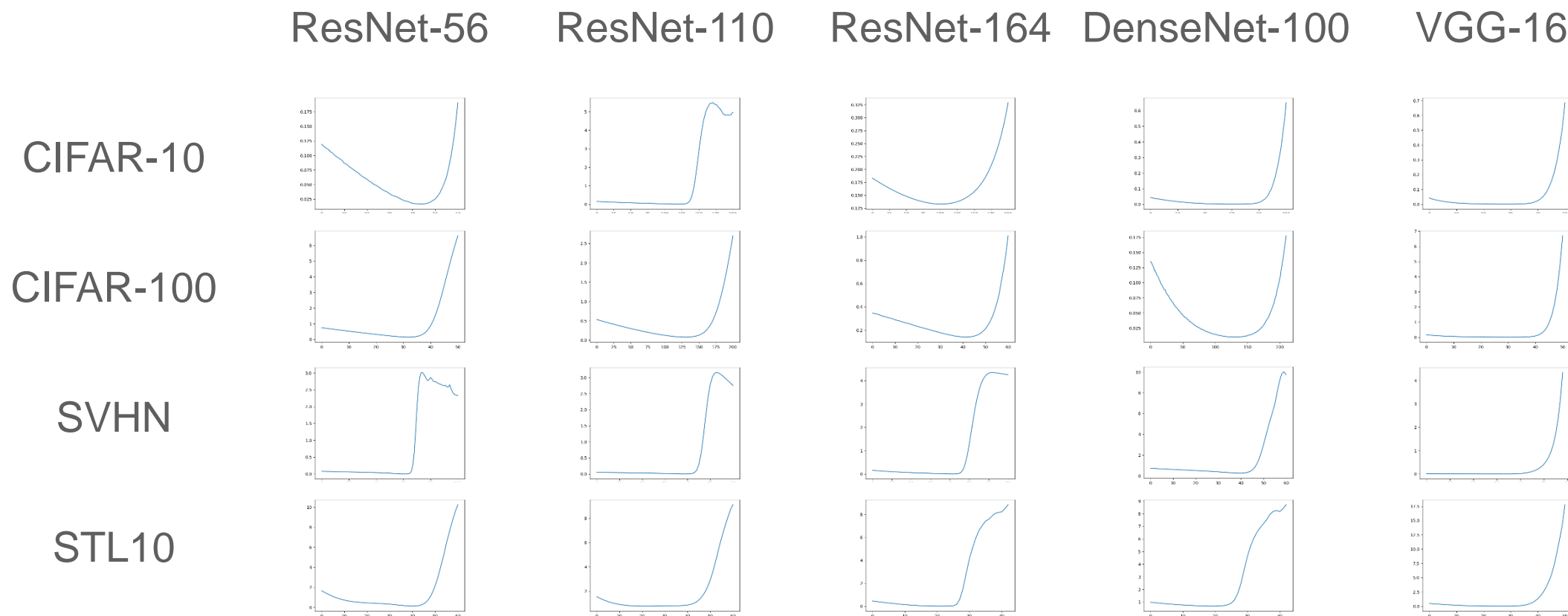
- A direction  $u$  is  $(\bar{r}, \underline{r}, p, c)$ -asymmetric with respect to  $w$  if  $\nabla_l \hat{L}(w + lu) < p$ ,  $\nabla_l \hat{L}(w - lu) > cp$  and  $l \in (\bar{r}, \underline{r})$



# Our Proposal: Asymmetric Valley



Wide existence of asymmetric direction

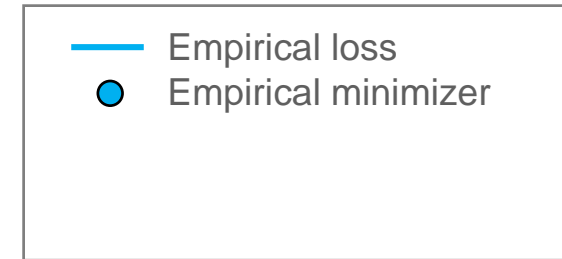
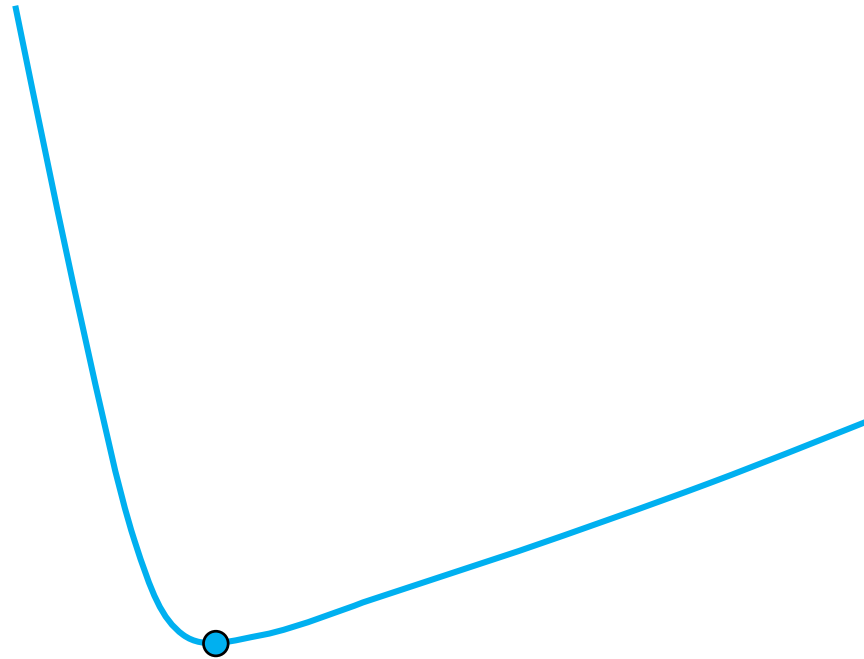




# Asymmetric Valley and Generalization

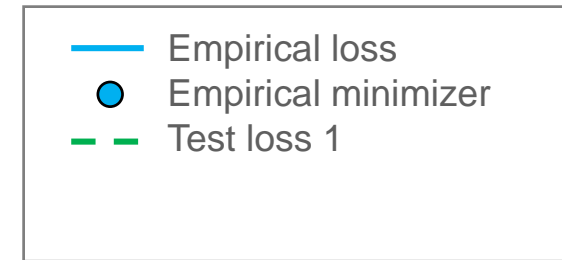
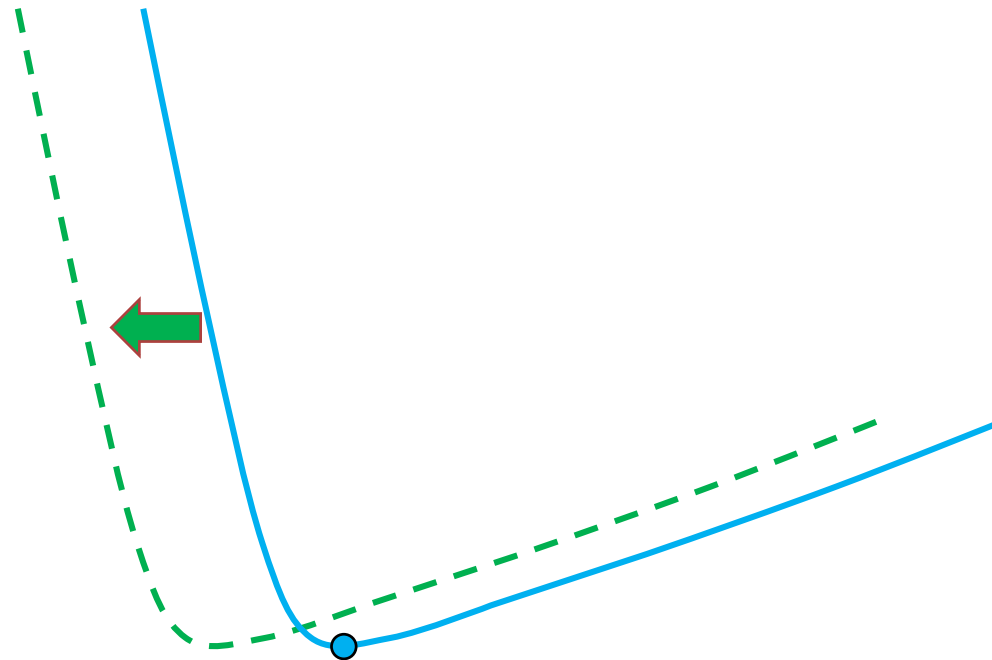


## Case I: Empirical minimizer



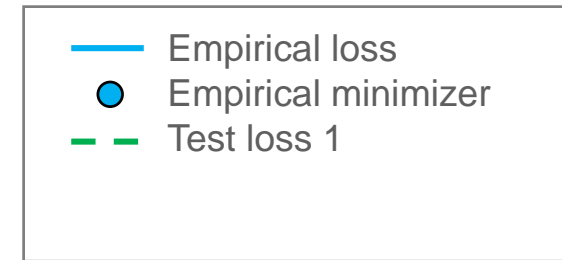
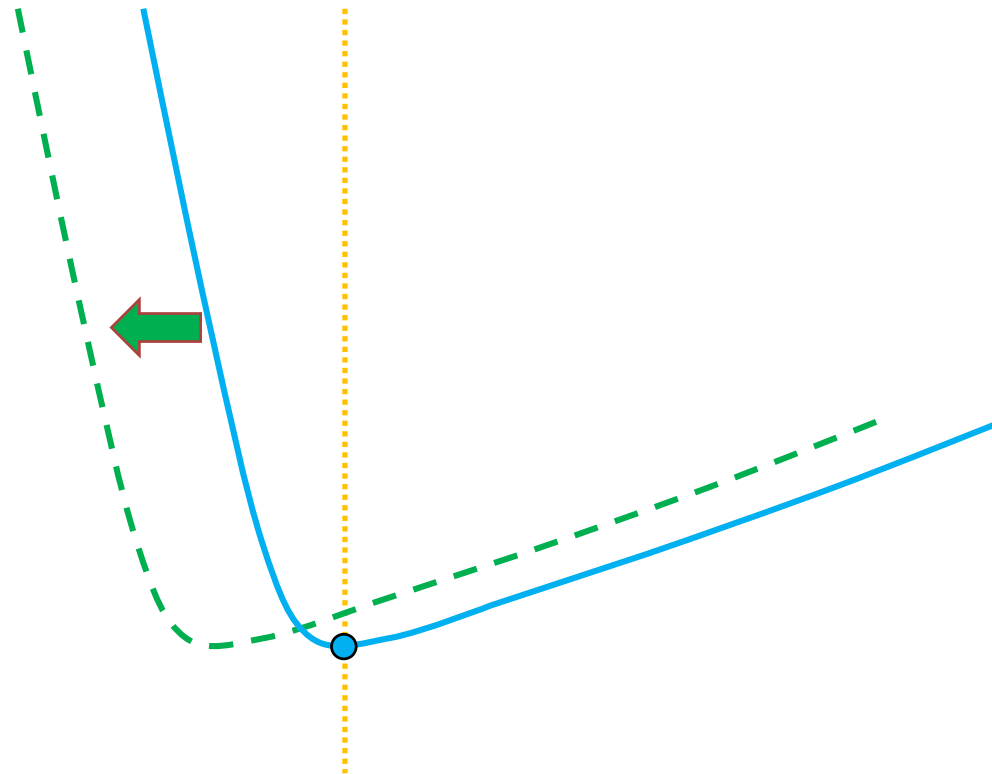


## Case I: Empirical minimizer





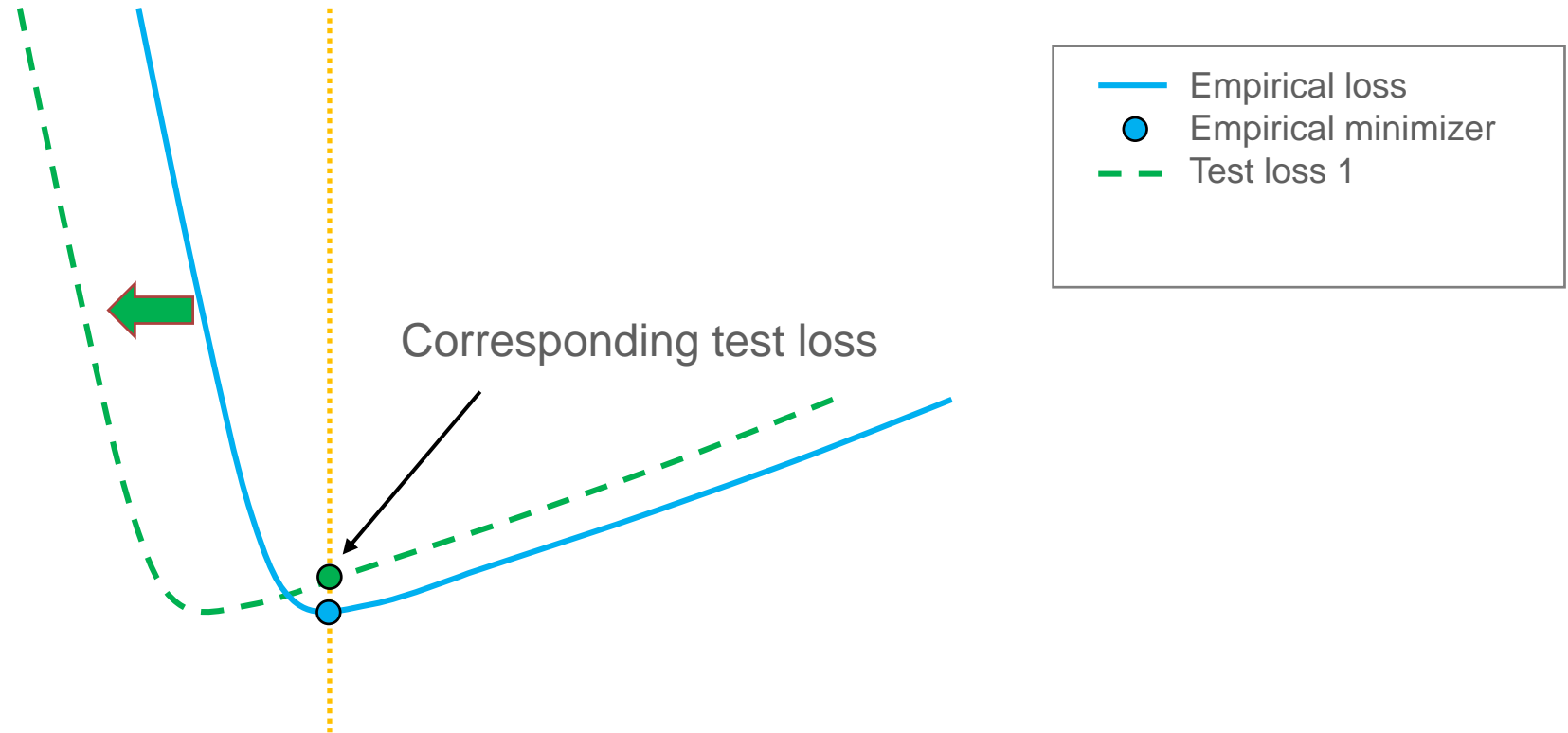
## Case I: Empirical minimizer



# Our Proposal: Asymmetric Valley

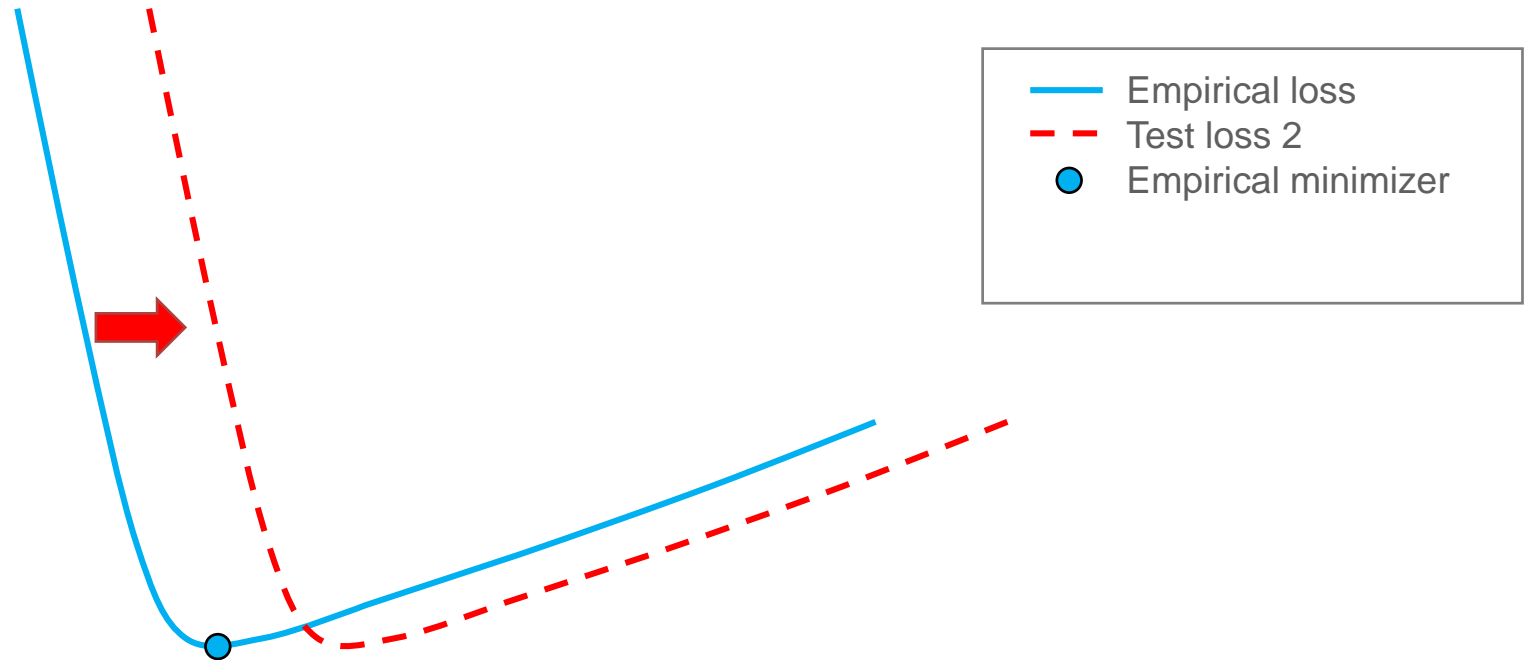


## Case I: Empirical minimizer





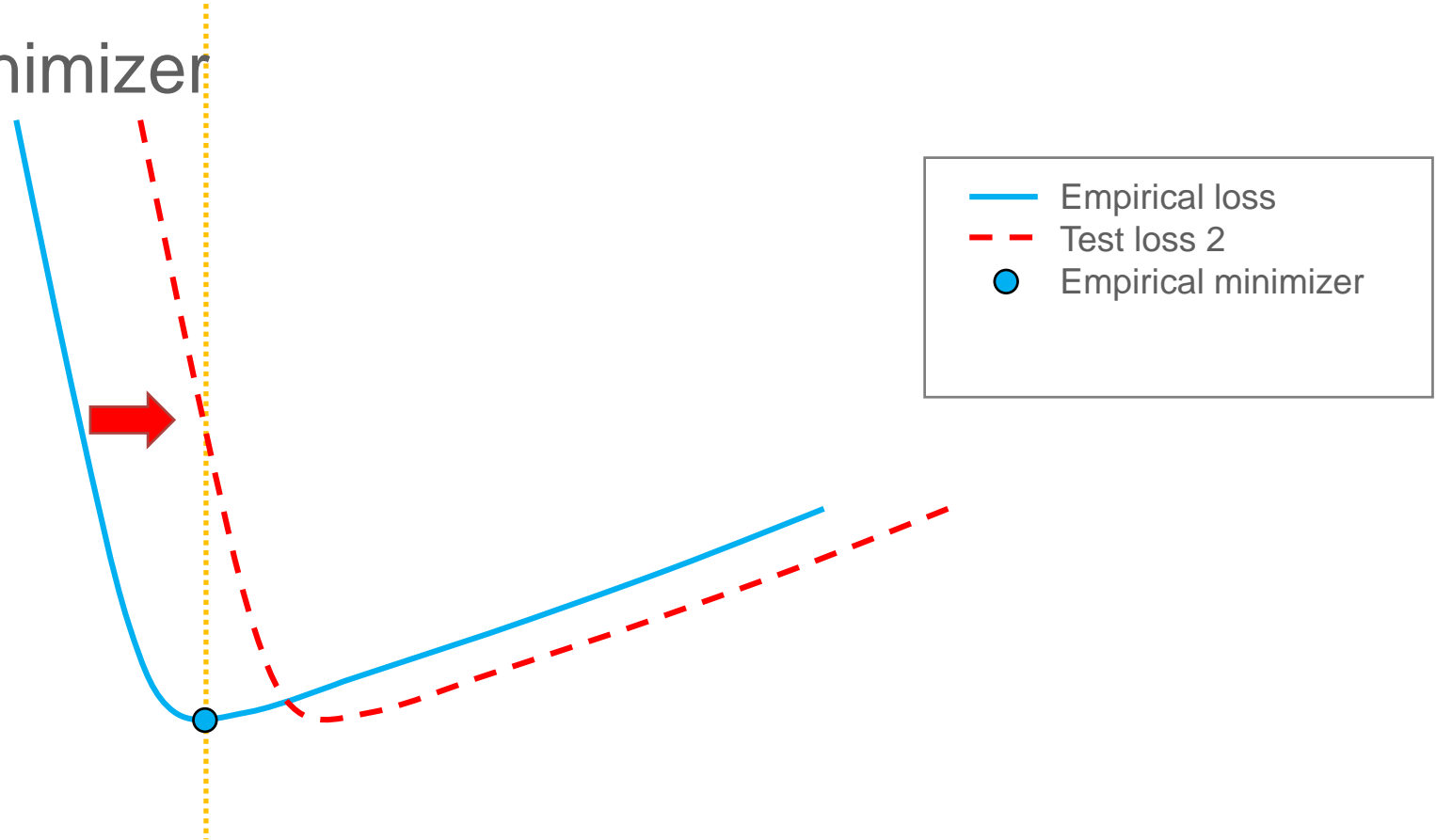
## Case I: Empirical minimizer



# Asymmetric Valley and Generalization



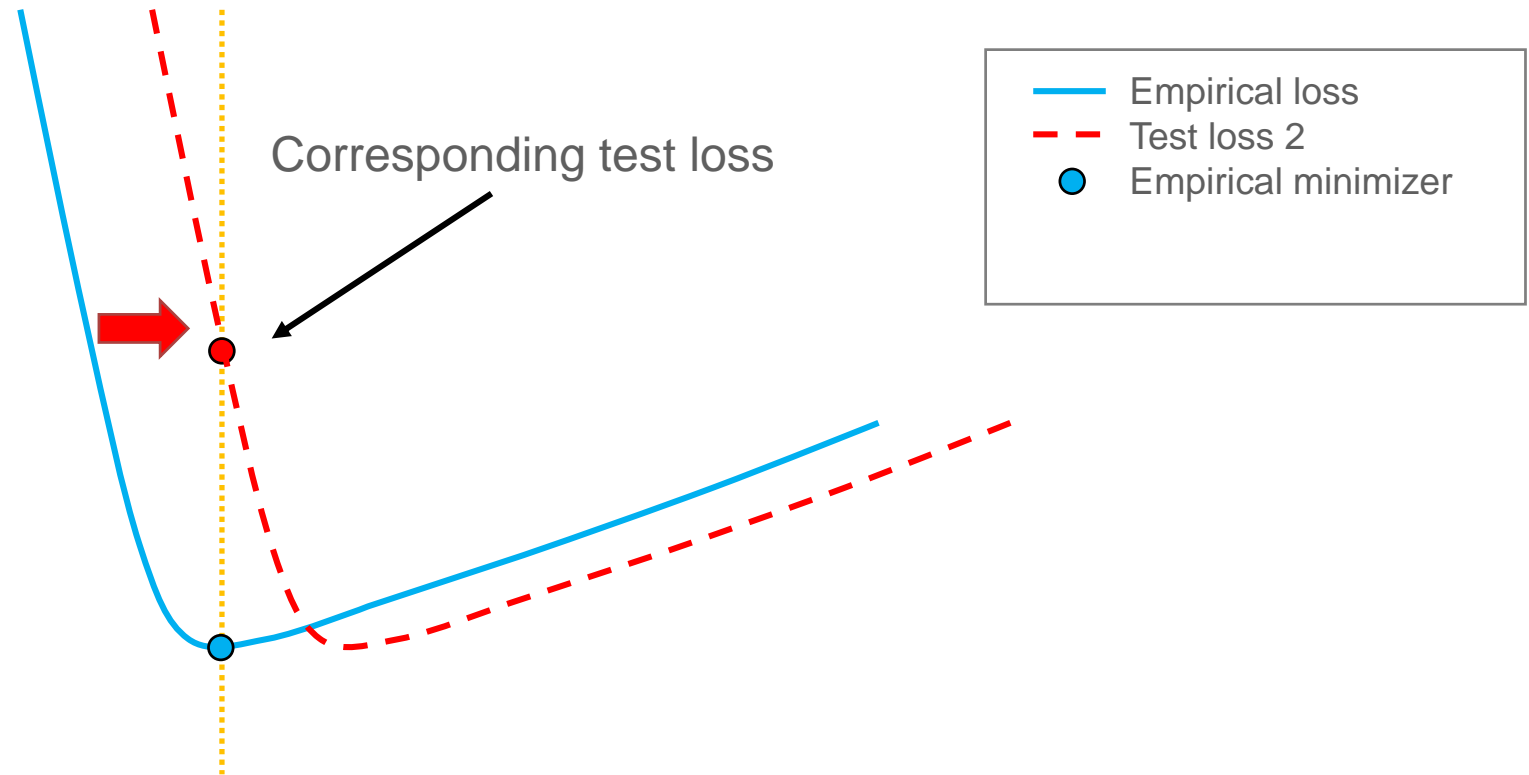
## Case I: Empirical minimizer



# Asymmetric Valley and Generalization



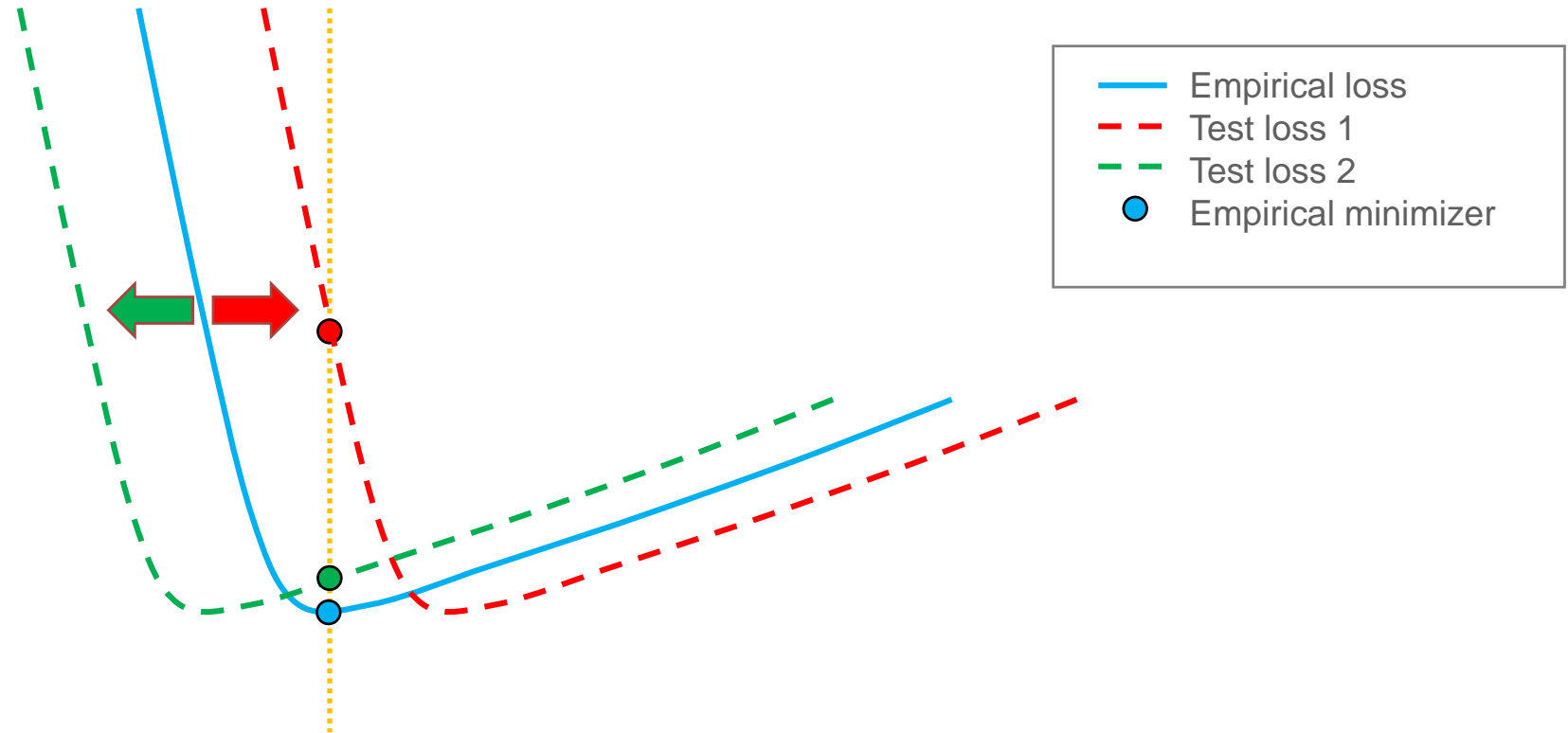
## Case I: Empirical minimizer



# Asymmetric Valley and Generalization



## Case I: Empirical minimizer

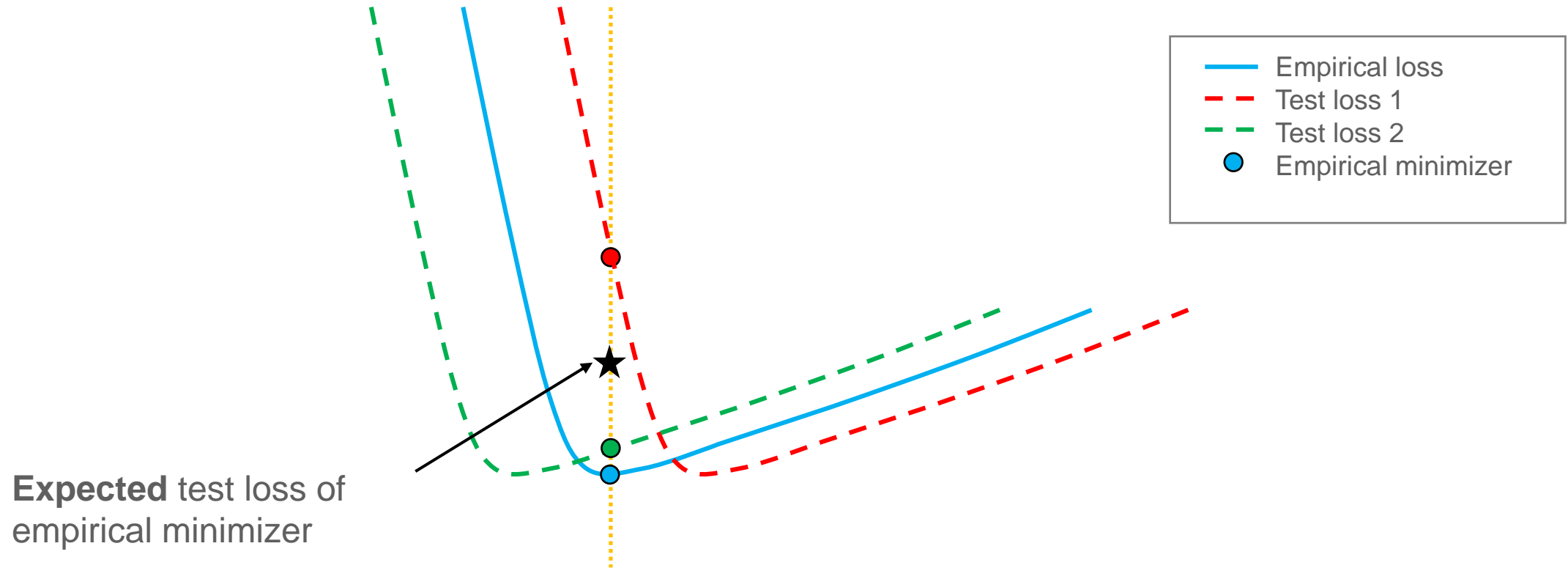




# Asymmetric Valley and Generalization



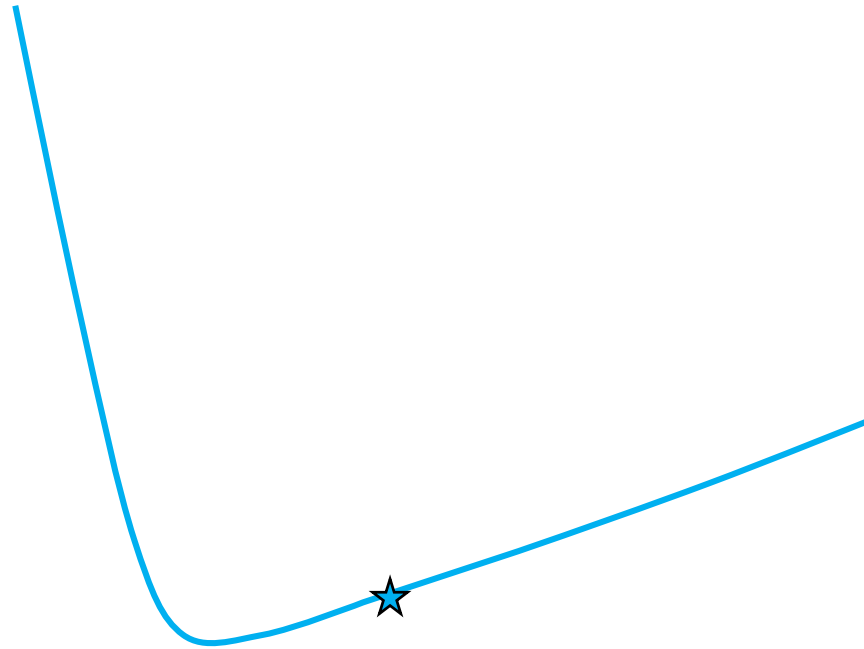
## Case I: Empirical minimizer



# Asymmetric Valley and Generalization



Case II: Biased solution towards the **flat** side

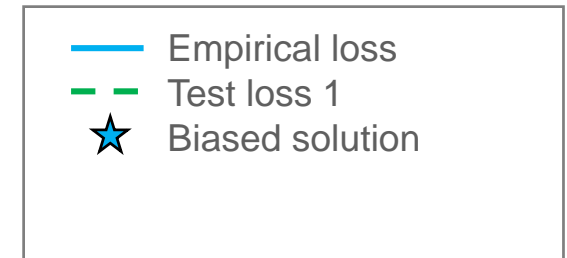
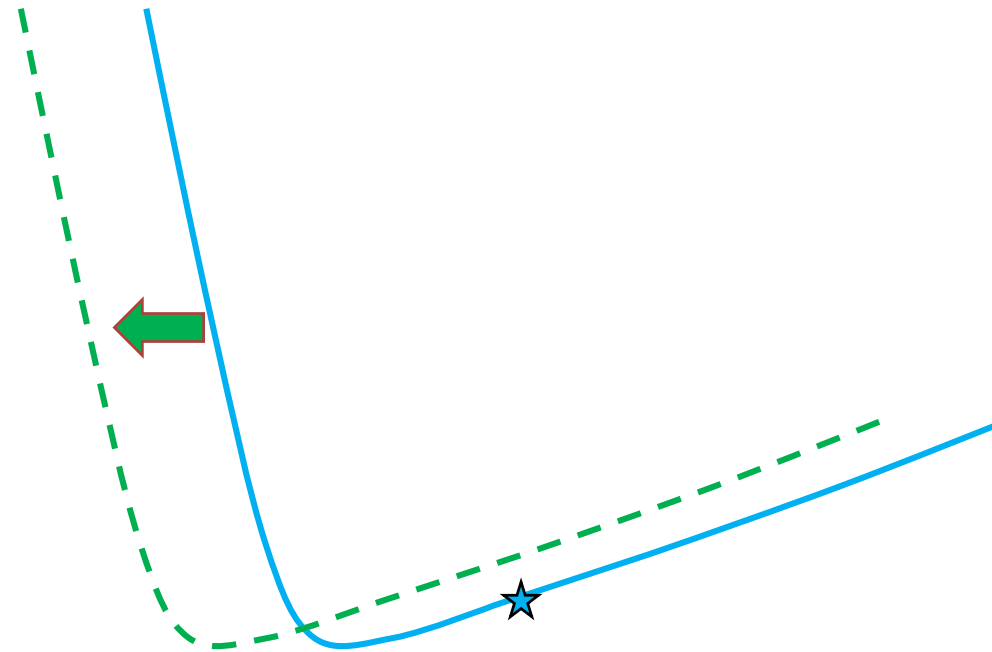


— Empirical loss  
★ Biased solution

# Asymmetric Valley and Generalization



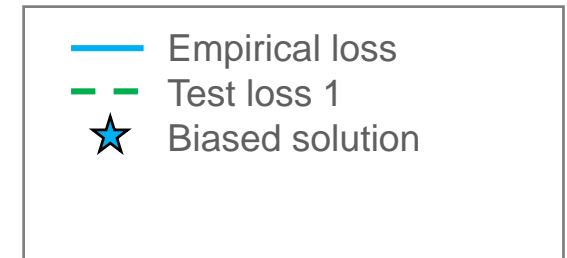
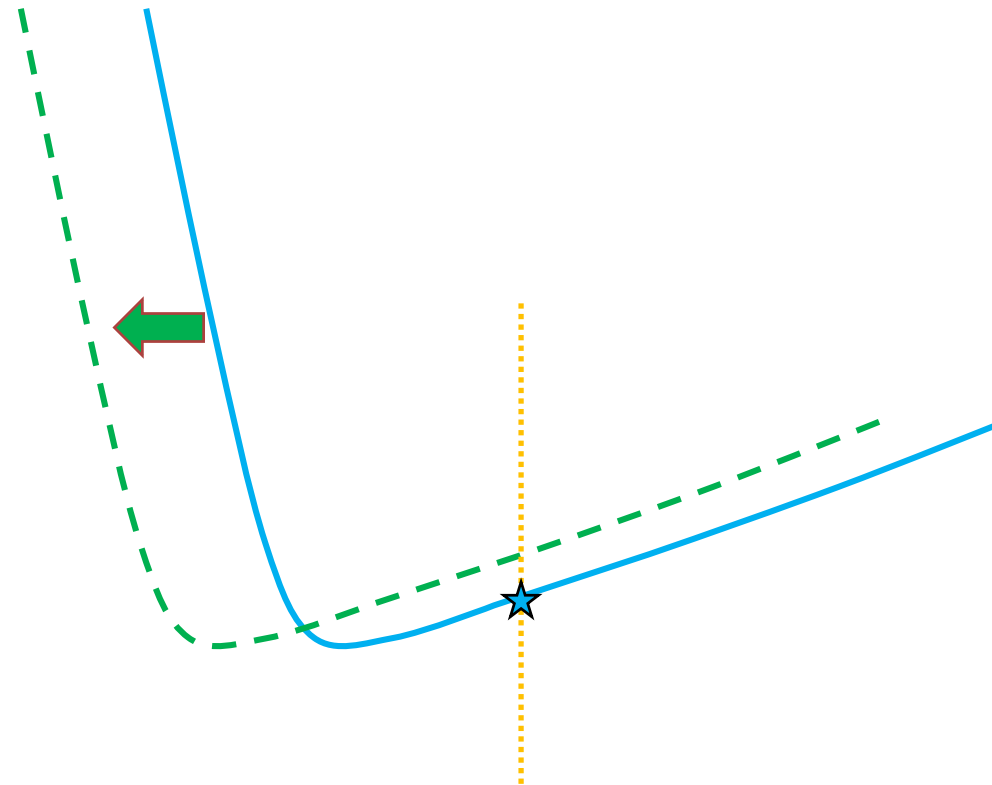
Case II: Biased solution towards the **flat** side



# Asymmetric Valley and Generalization



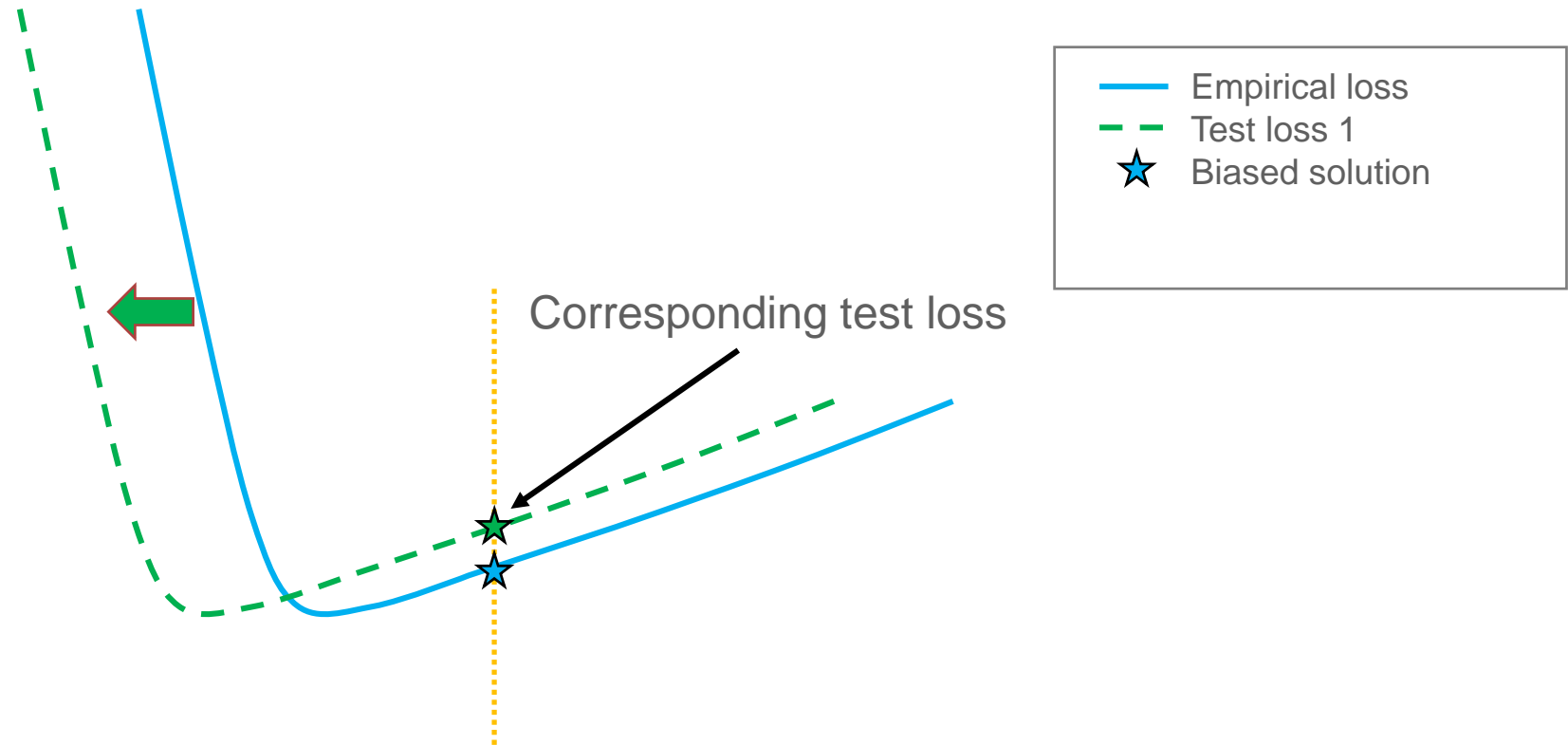
## Case II: Biased solution towards the **flat** side



# Asymmetric Valley and Generalization



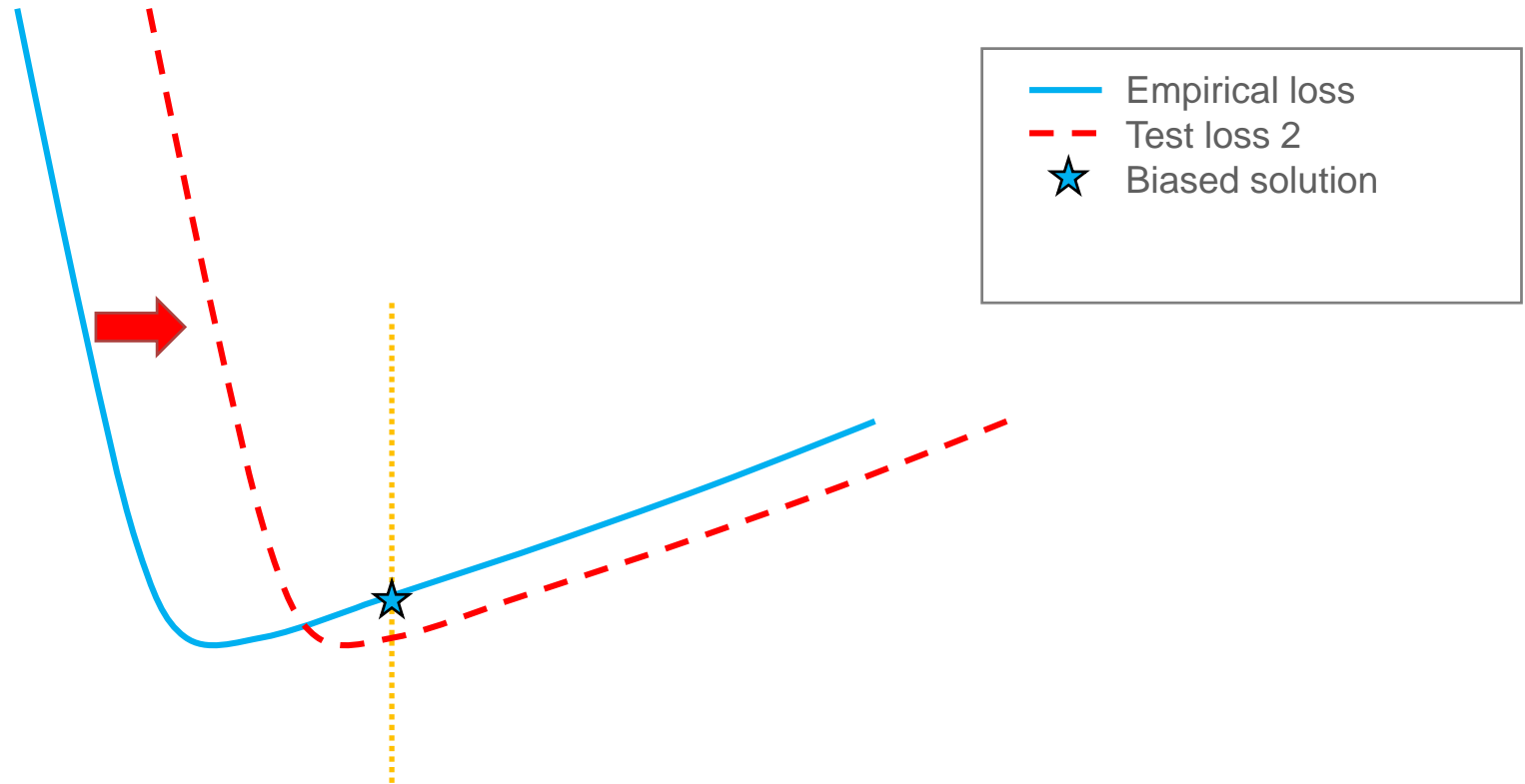
## Case II: Biased solution towards the flat side



# Asymmetric Valley and Generalization



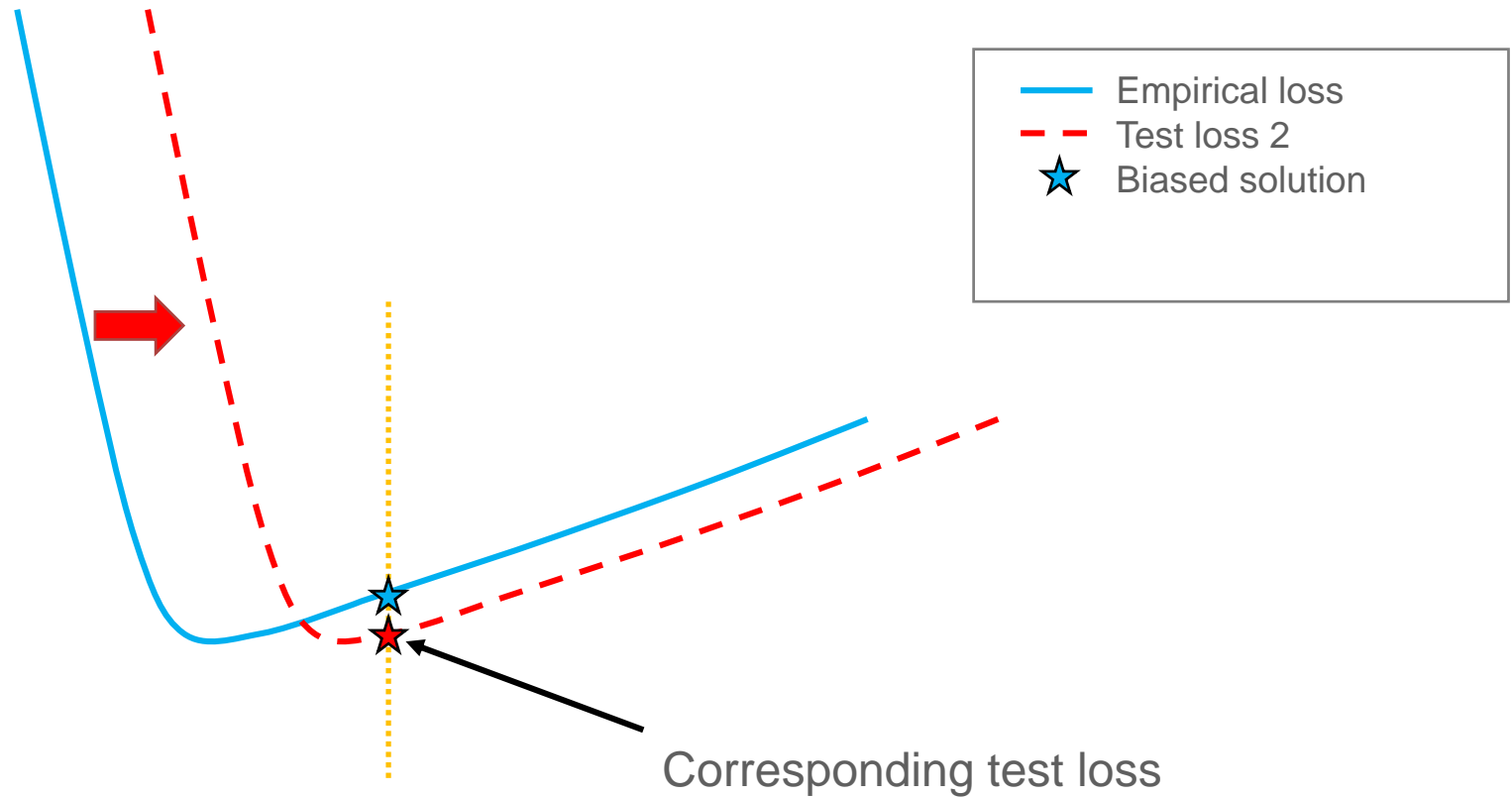
## Case II: Biased solution towards the **flat** side



# Asymmetric Valley and Generalization



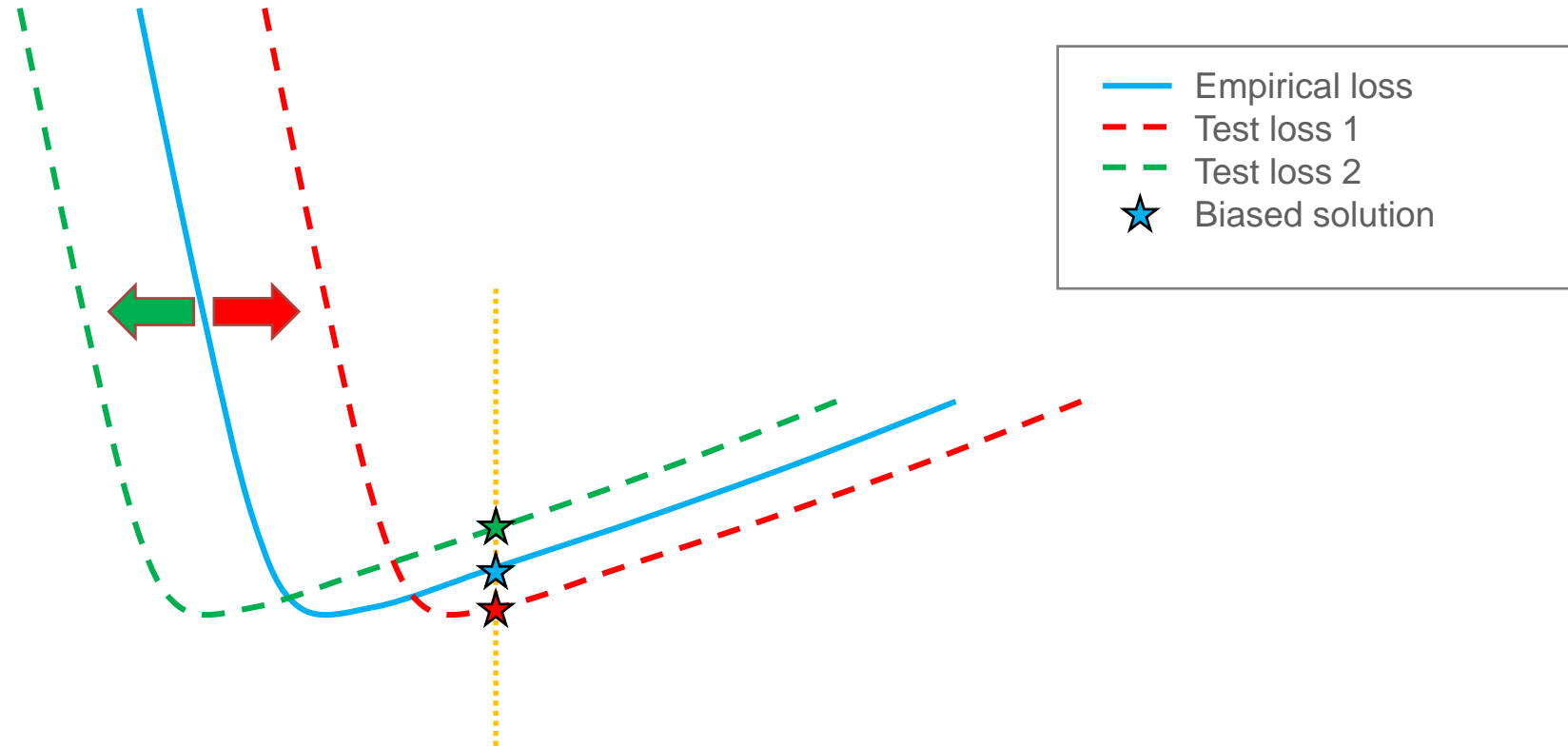
## Case II: Biased solution towards the **flat** side



# Asymmetric Valley and Generalization

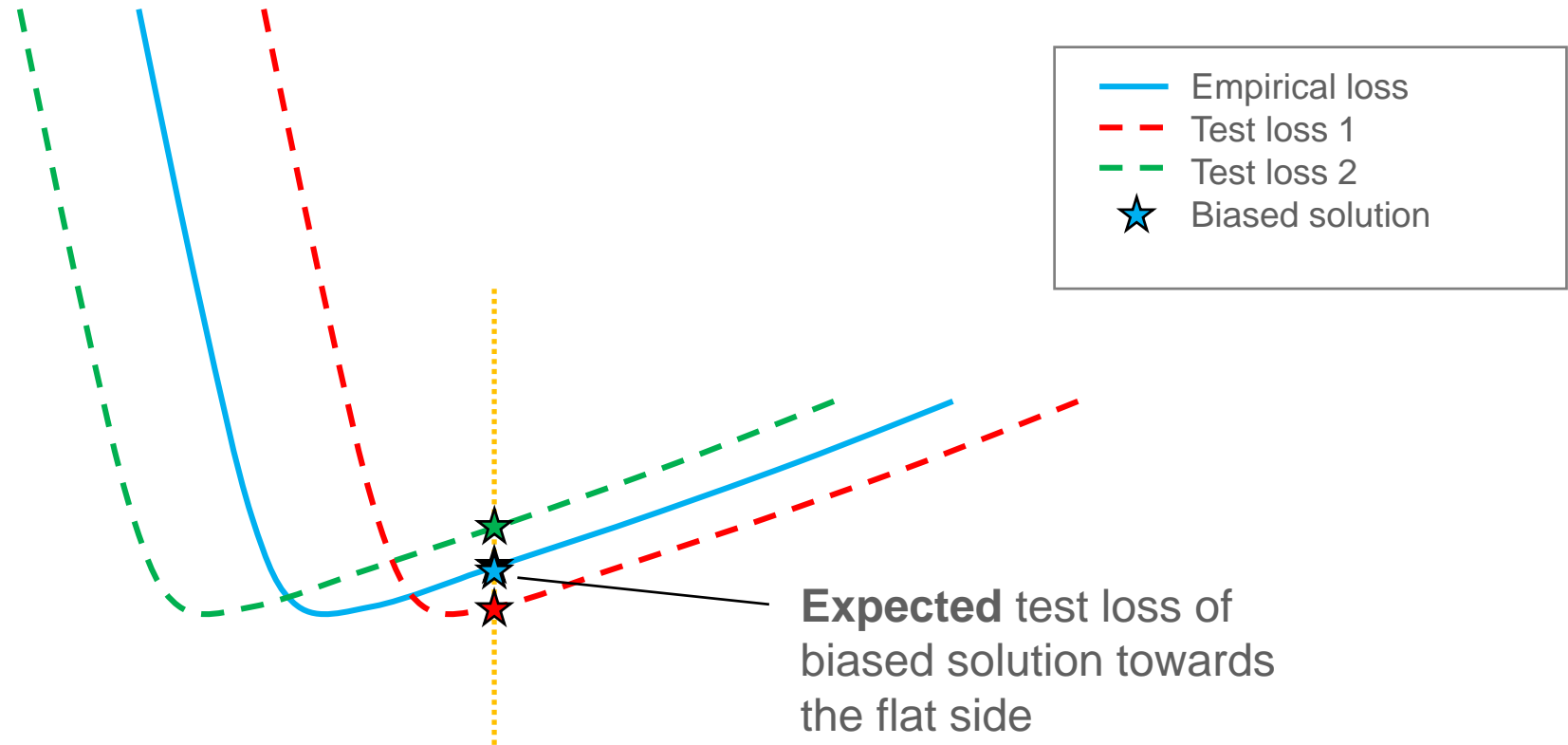


## Case II: Biased solution towards the flat side





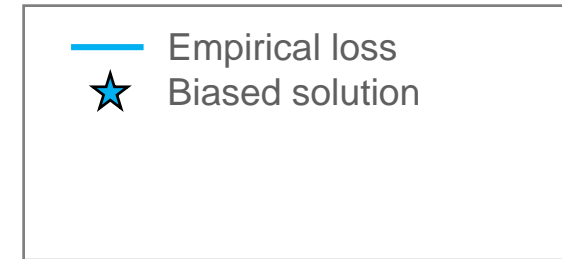
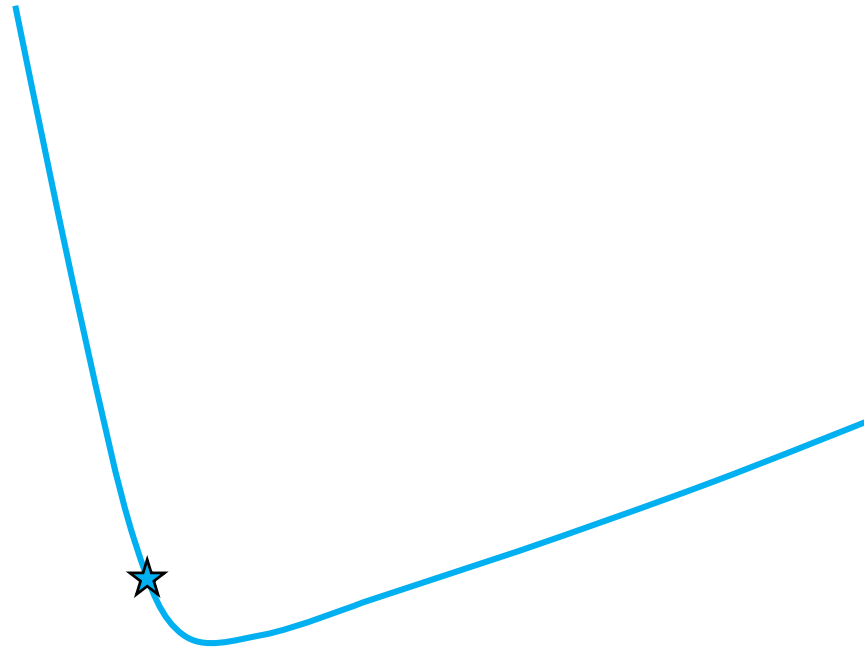
## Case II: Biased solution towards the **flat** side



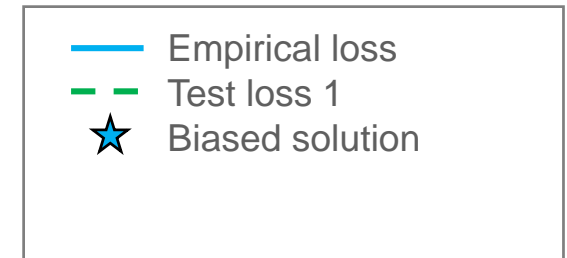
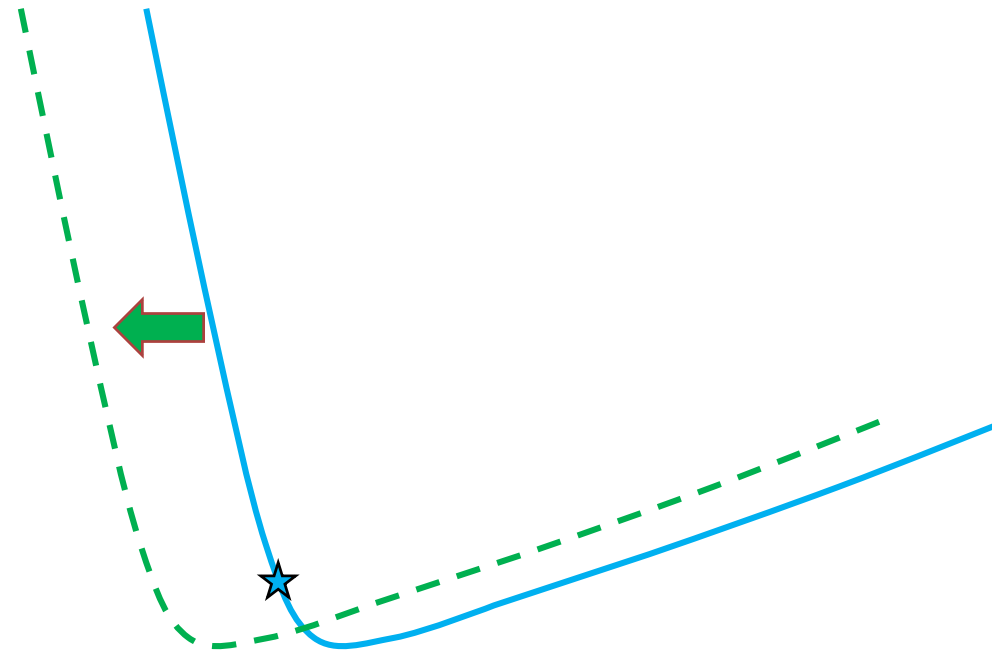
# Asymmetric Valley and Generalization



## Case III: Biased solution towards the sharp side



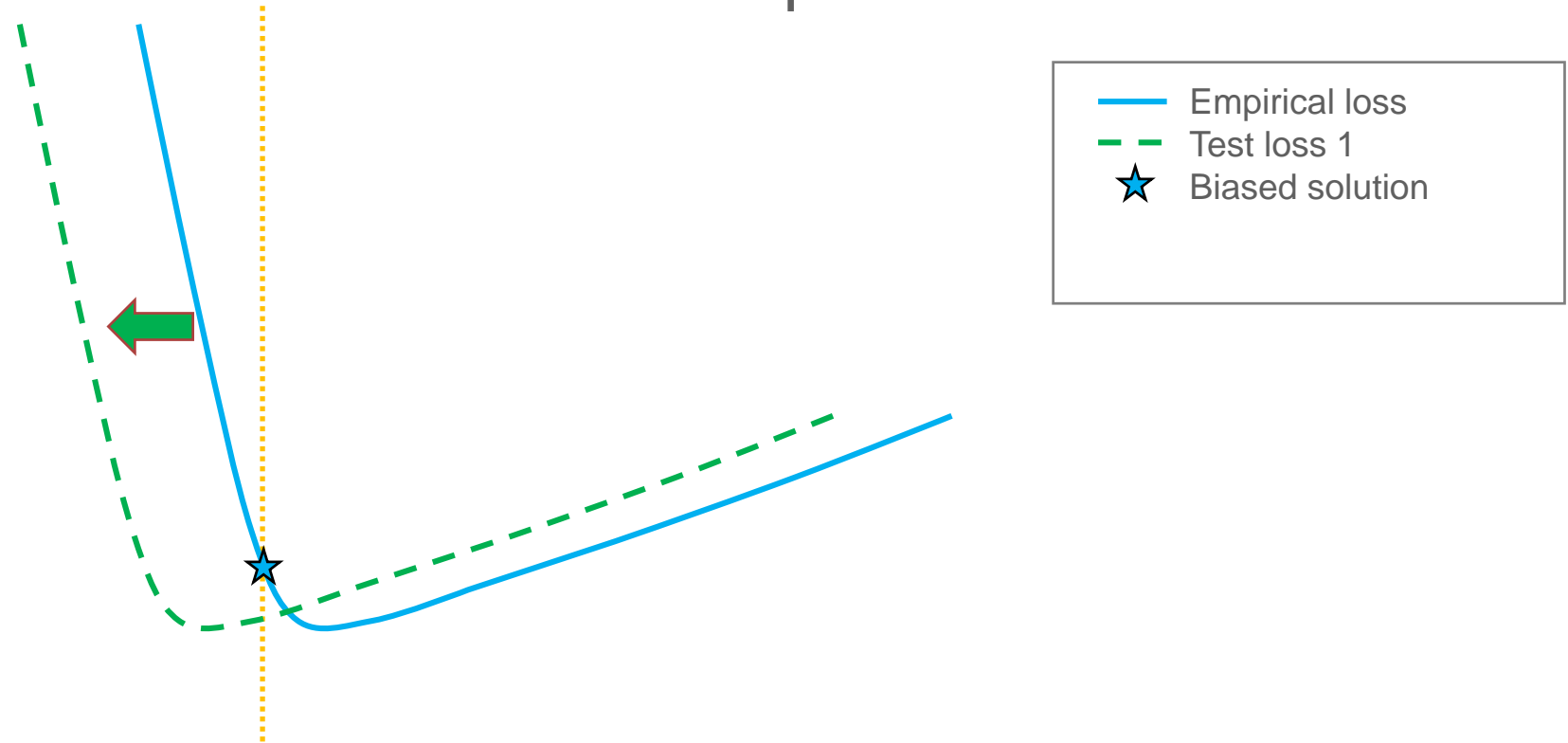
## Case III: Biased solution towards the sharp side



# Asymmetric Valley and Generalization



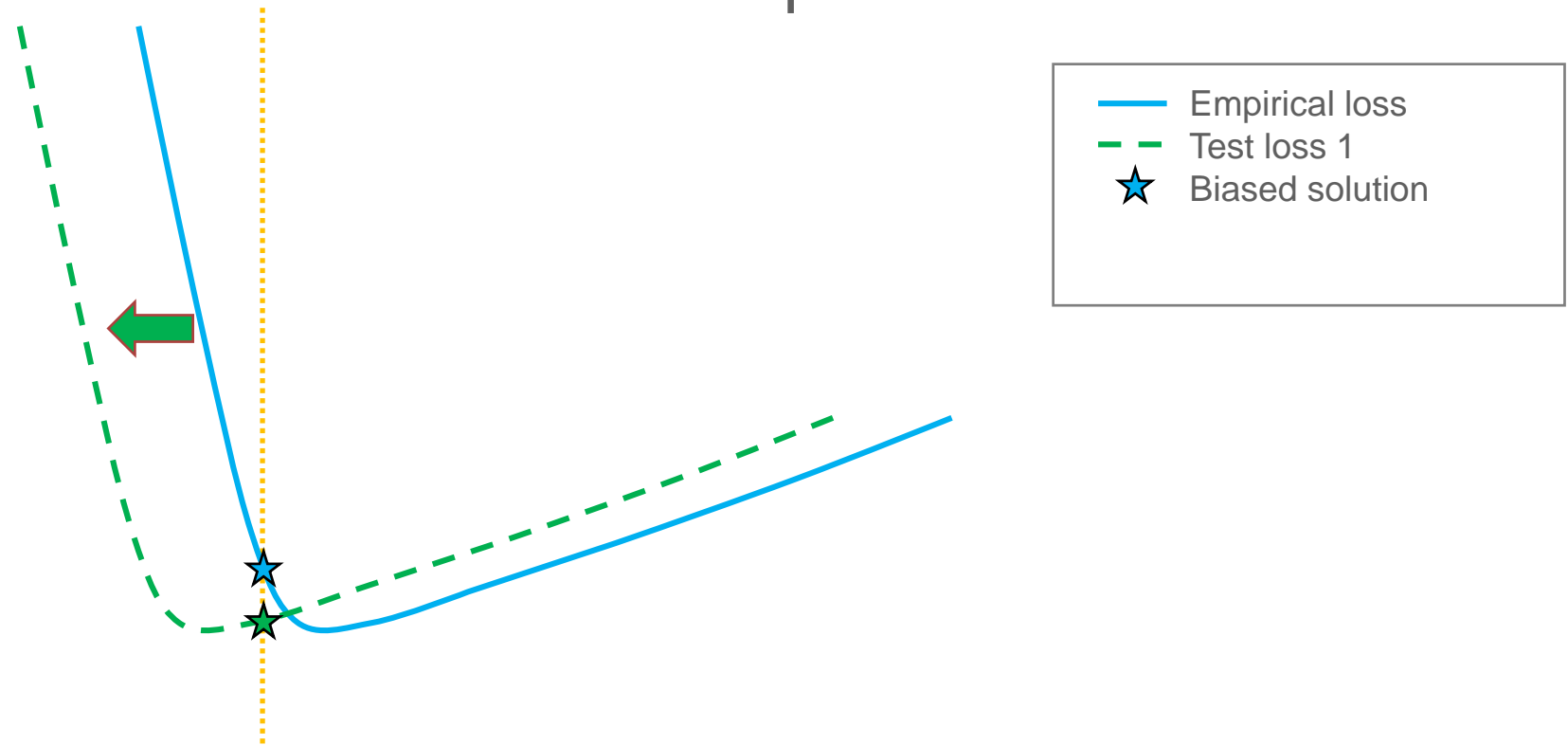
## Case III: Biased solution towards the sharp side



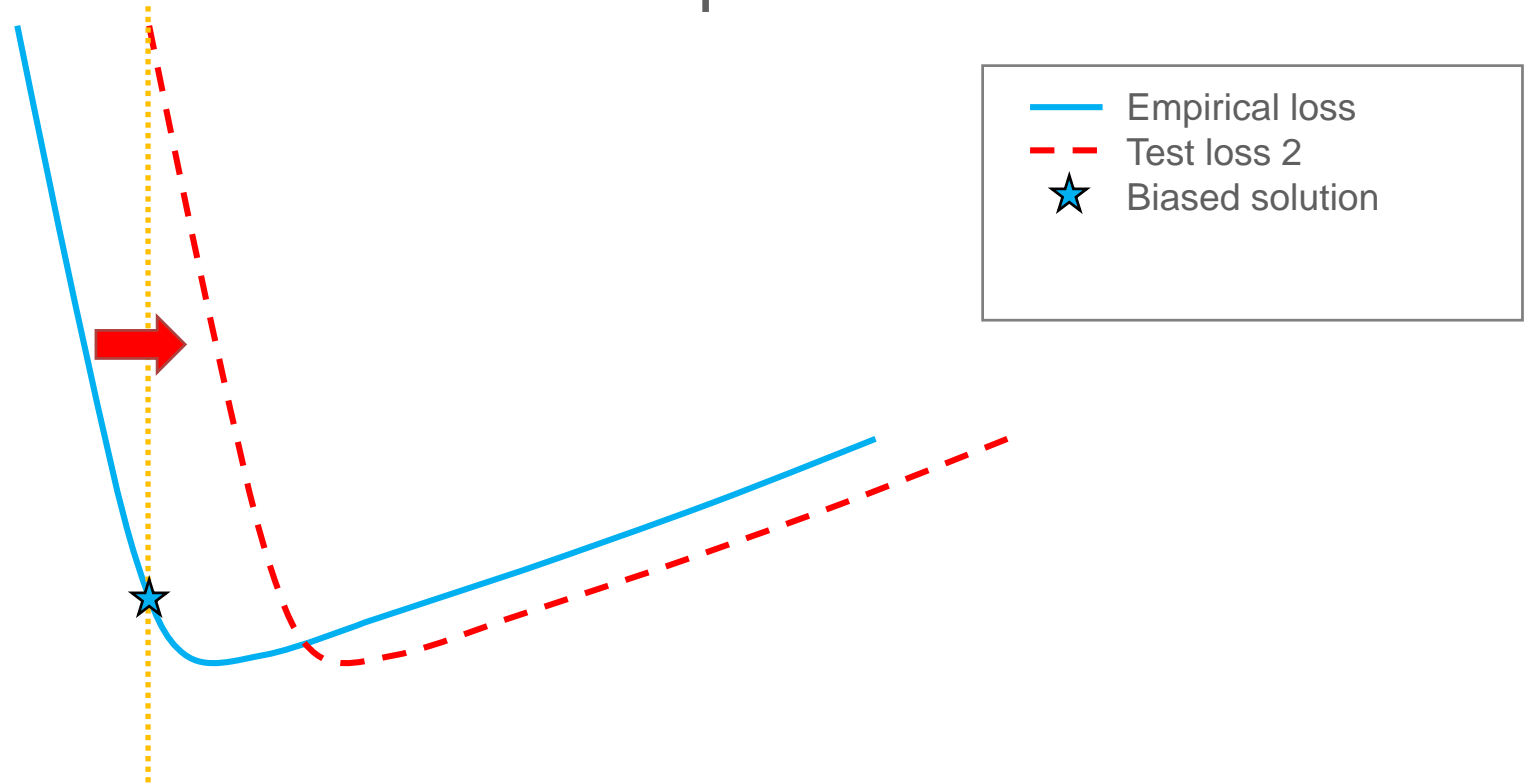
# Asymmetric Valley and Generalization



## Case III: Biased solution towards the sharp side



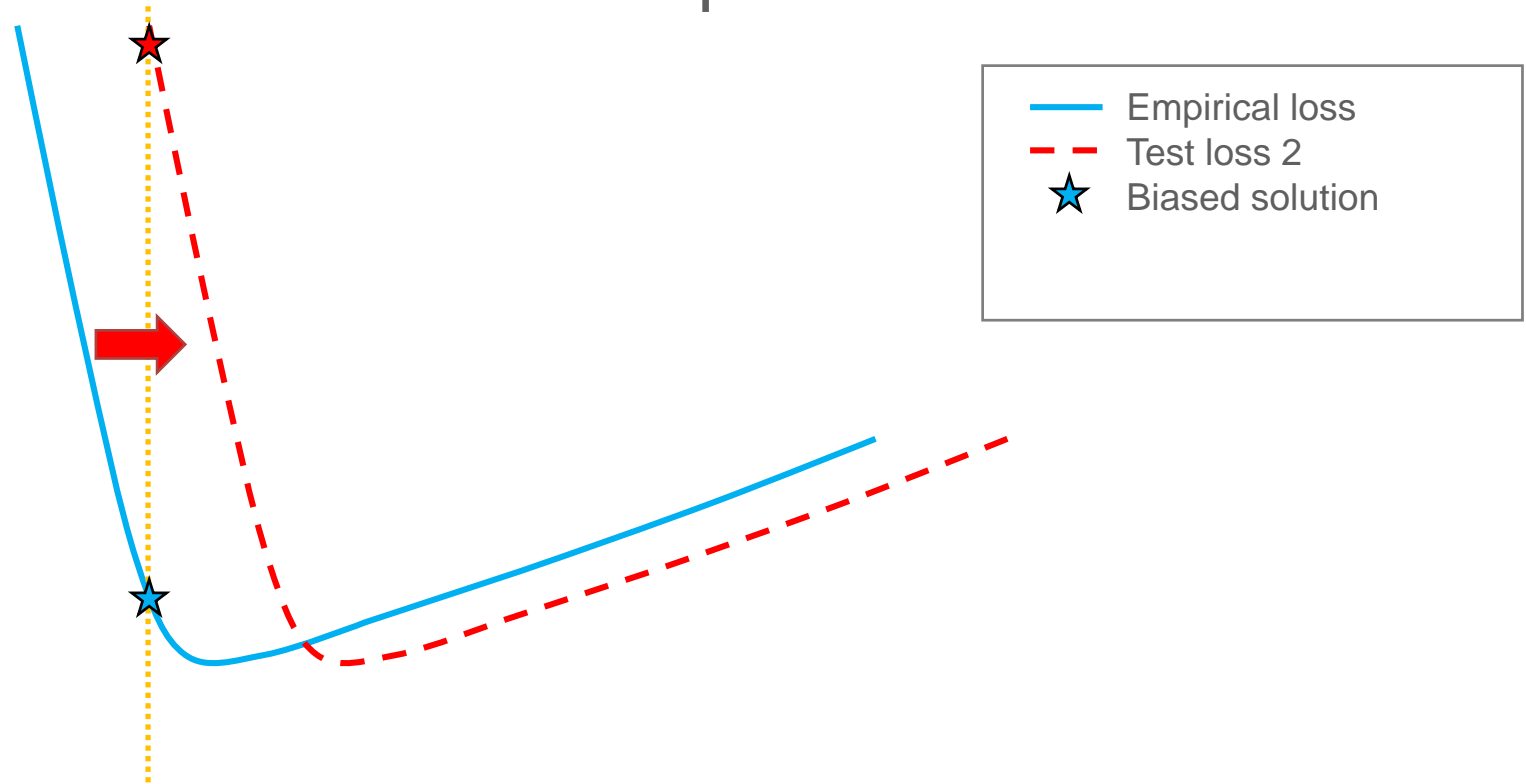
## Case III: Biased solution towards the sharp side



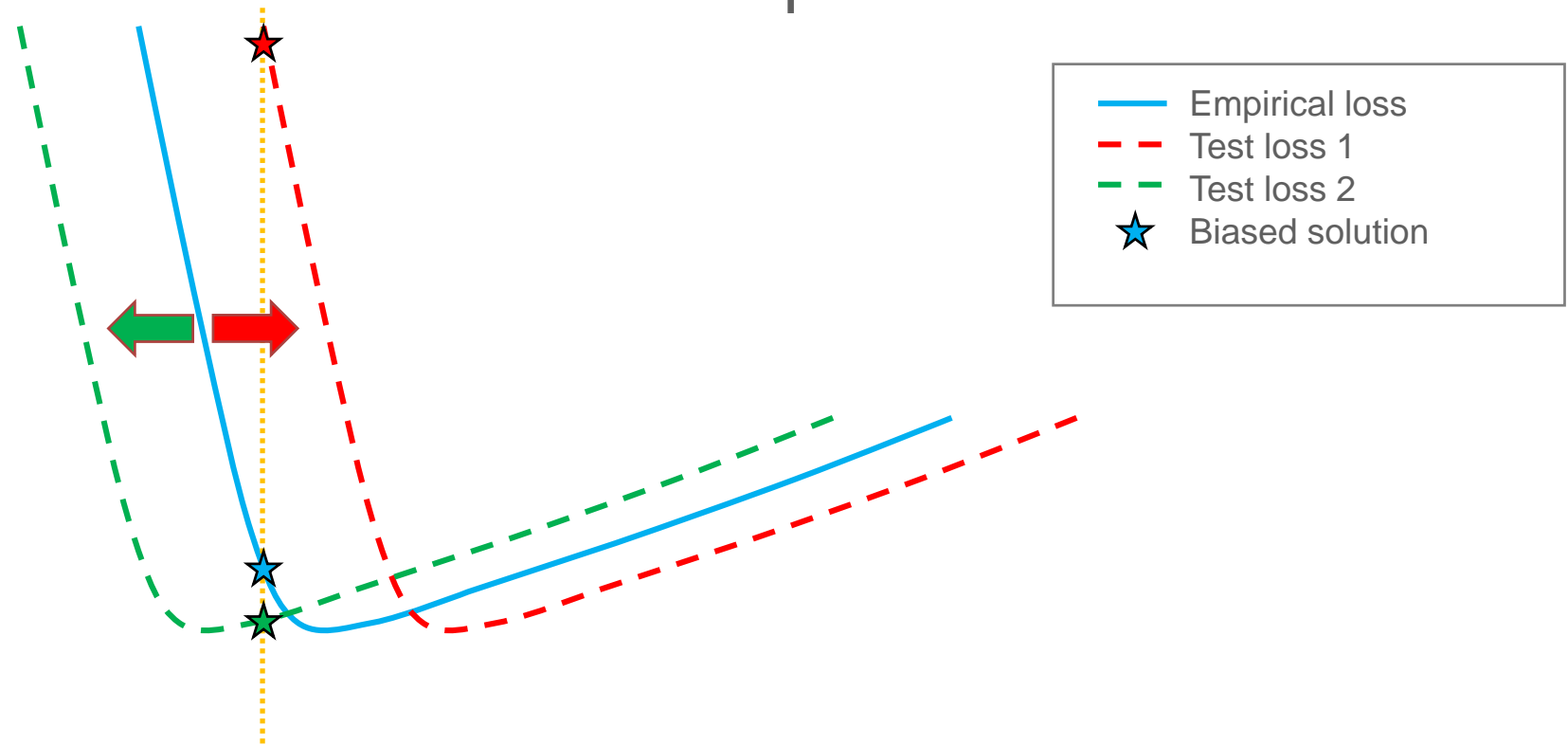
# Asymmetric Valley and Generalization



## Case III: Biased solution towards the sharp side



## Case III: Biased solution towards the sharp side

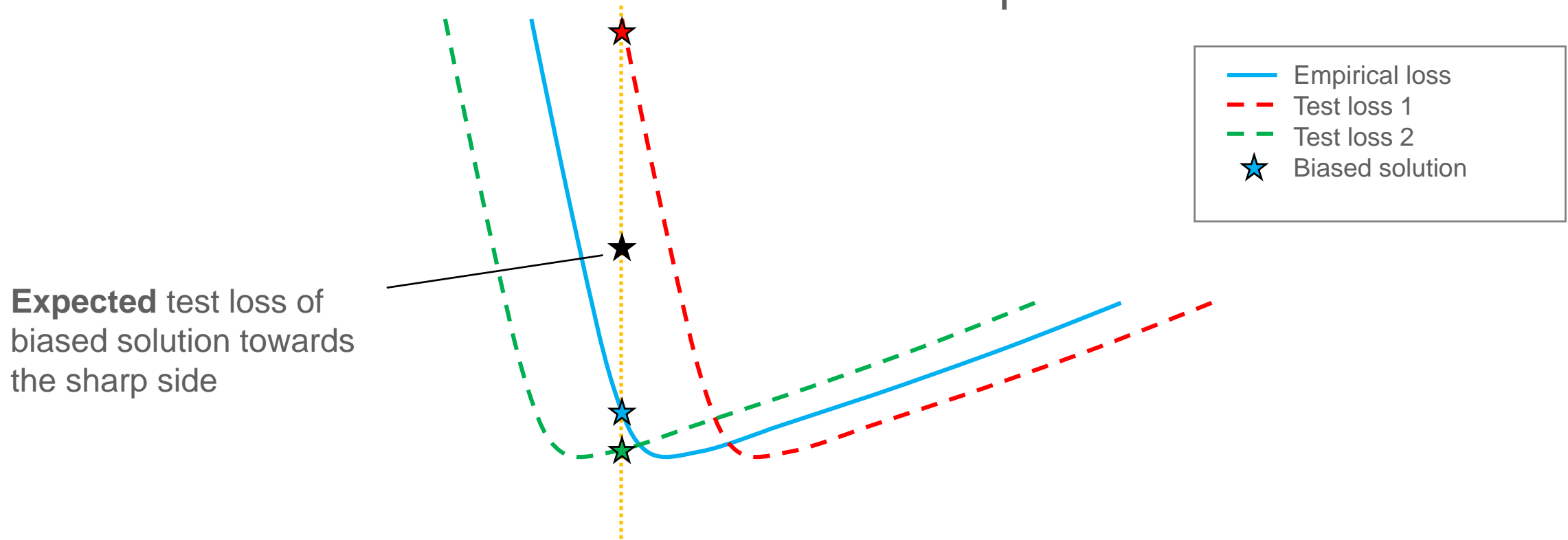




# Asymmetric Valley and Generalization



## Case III: Biased solution towards the sharp side



# Our Proposal: Asymmetric Valley



清华大学  
Tsinghua University



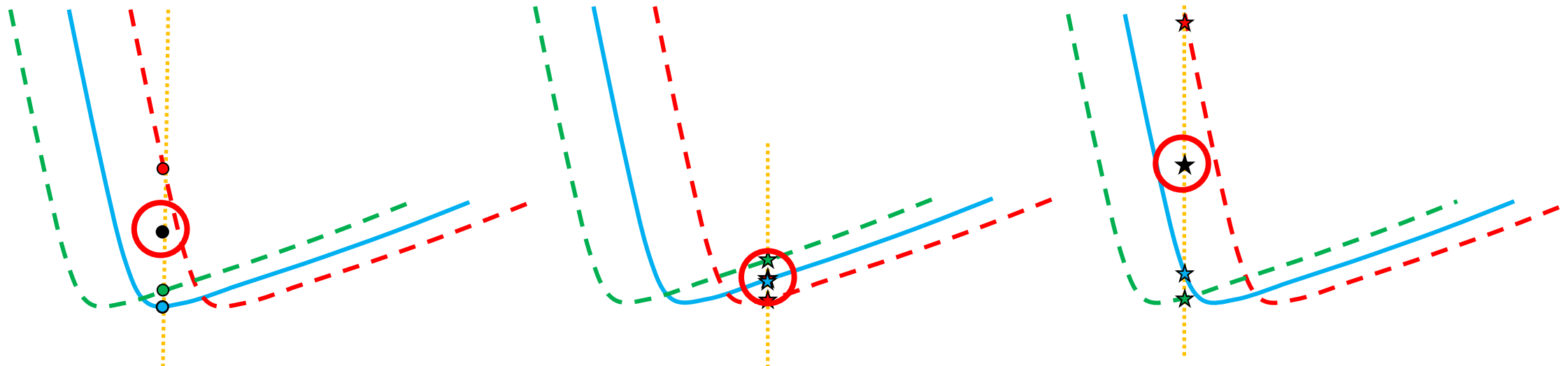
Case I:  
Empirical Minimizer



Case II:  
Biased towards the **flat** side



Case III:  
Biased towards the **sharp** side



**Flat** side biased solution (**Case II**) generalize better!

# Main Theorem 1



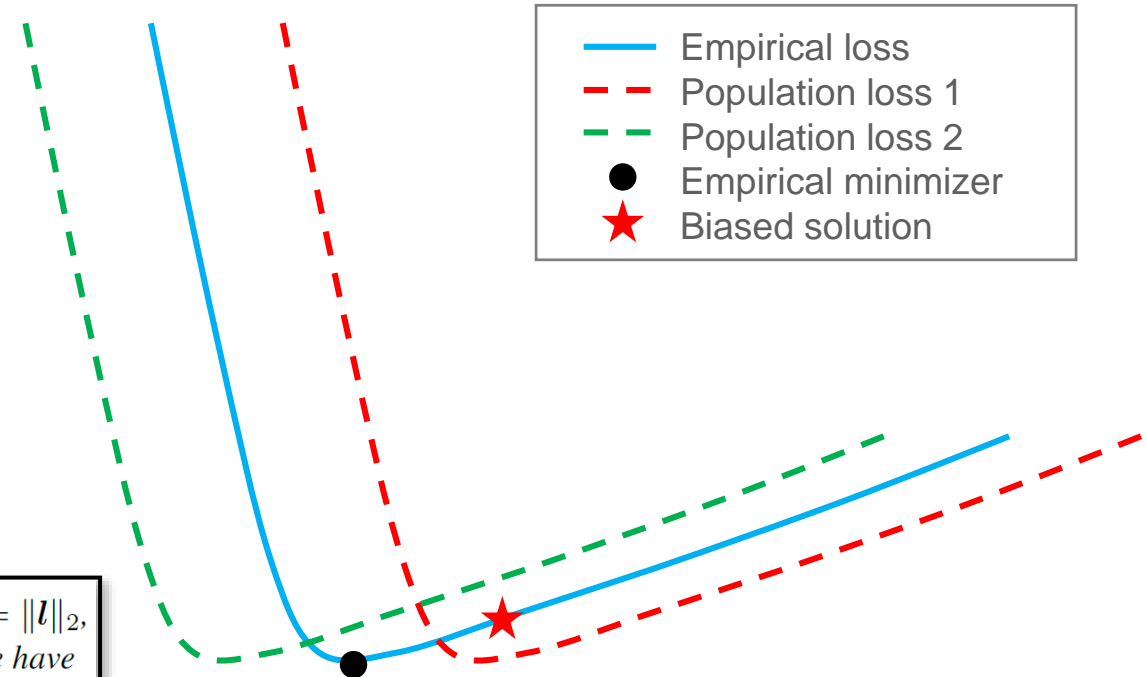
Biased solution on the flat side of an asymmetric valley leads to better generalization

$$E_{\delta}L(\hat{w}^*) - E_{\delta}L(\hat{w}^* + c_0) > 0$$

where  $c_0$  is a bias towards the flat side,  $\hat{w}^*$  is an empirical solution

**Theorem 1** (Bias leads to better generalization). For any  $\mathbf{l} \in \mathbb{R}^k$ , if Assumption 1 holds for  $R = \|\mathbf{l}\|_2$ , Assumption 2 holds for  $R' = \|\bar{\delta}\|_2 + \|\mathbf{l}\|_2$ , and  $\frac{4\xi}{(c_i-1)p_i} < l_i \leq \max\{\bar{r} - \bar{\delta}_i, \bar{\delta}_i - \underline{r}\}$ , then we have

$$\mathbb{E}_{\delta}L(\hat{w}^*) - \mathbb{E}_{\delta}L\left(\hat{w}^* + \sum_{i=1}^k l_i \mathbf{u}^i\right) \geq \sum_{i=1}^k (c_i - 1)l_i p_i / 2 - 2k\xi > 0$$

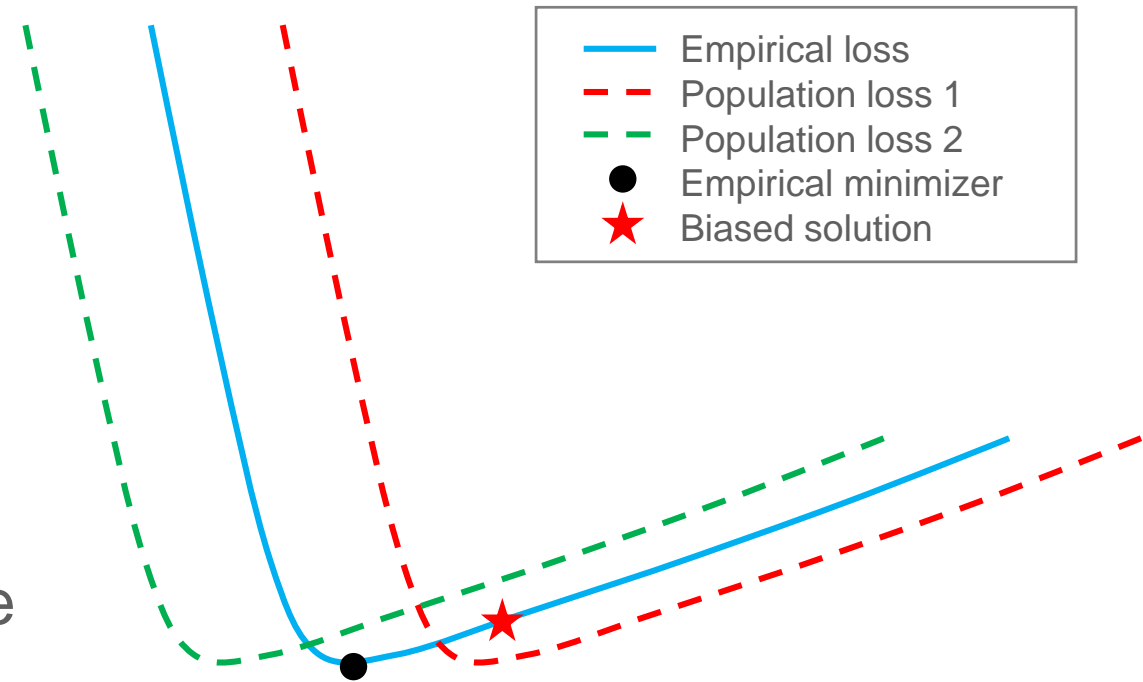


# Main Theorem 1



## Two interesting implications:

- Converging to *which* local minimum may not be critical. However, it matters *where* the solution locates in a basin.
- The solution with **lowest generalization** error is **not** necessarily the minimizer of the training loss.





How to obtain a biased solution towards the flat side of an asymmetric valley, empirically?

# Main Theorem 2 (informal)



Taking the **average** of the weights along the path of SGD leads to a biased solution towards the flat side

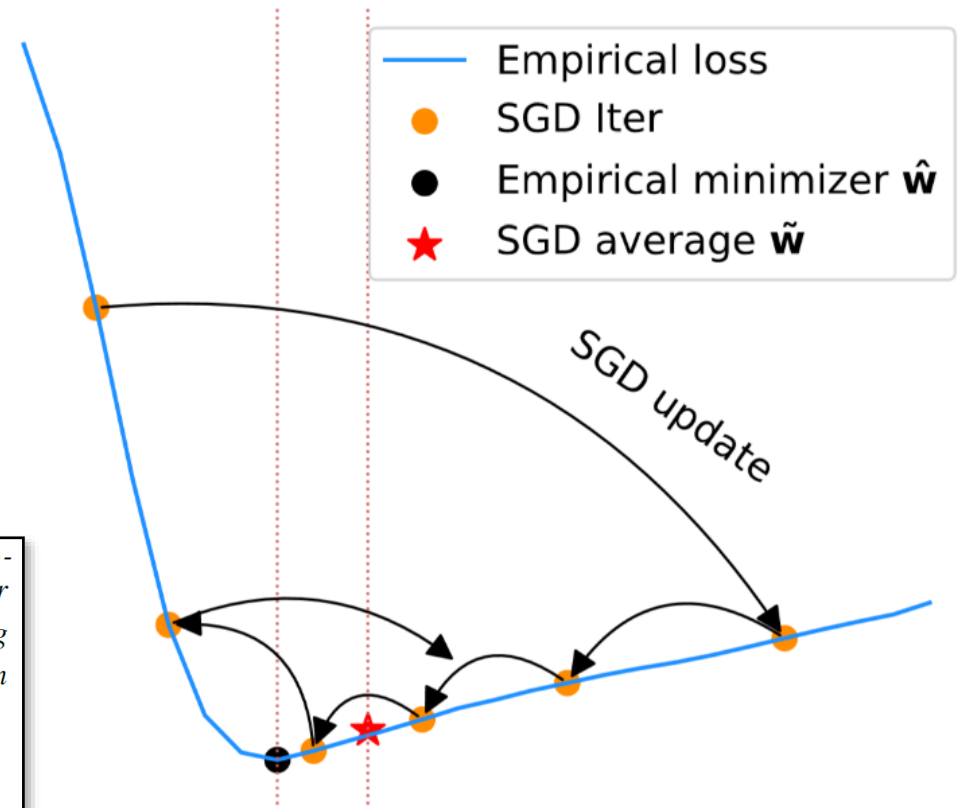
$$E[\bar{w}] > c_0 > 0$$

where  $c_0$  is a bias towards the flat side,  
 $\bar{w}$  is SGD average

**Theorem 2** (SGD averaging generates a bias). Assume that a local minimizer  $w^* = 0$  is a  $(r, 0, a_+, c)$ -asymmetric valley, where  $b_- \leq \nabla L(w) \leq a_- < 0$  for  $w < 0$ , and  $0 < b_+ \leq \nabla L(w) \leq a_+$  for  $w \geq 0$ . Assume  $-a_- = ca_+$  for a large constant  $c$ , and  $\frac{-(b_- - \nu)}{b_+} = c' < \frac{e^{c/3}}{6}$ . The SGD updating rule is  $w_{t+1} = w_t - \eta(\nabla L(w) + \omega_t)$  where  $\omega_t$  is the noise and  $|\omega_t| < \nu$ , and assume  $\nu \leq a_+$ . Then we have

$$E[\bar{w}] > c_0 > 0,$$

where  $c_0$  is a constant that only depends on  $\eta, a_+, a_-, b_+, b_-$  and  $\nu$ .

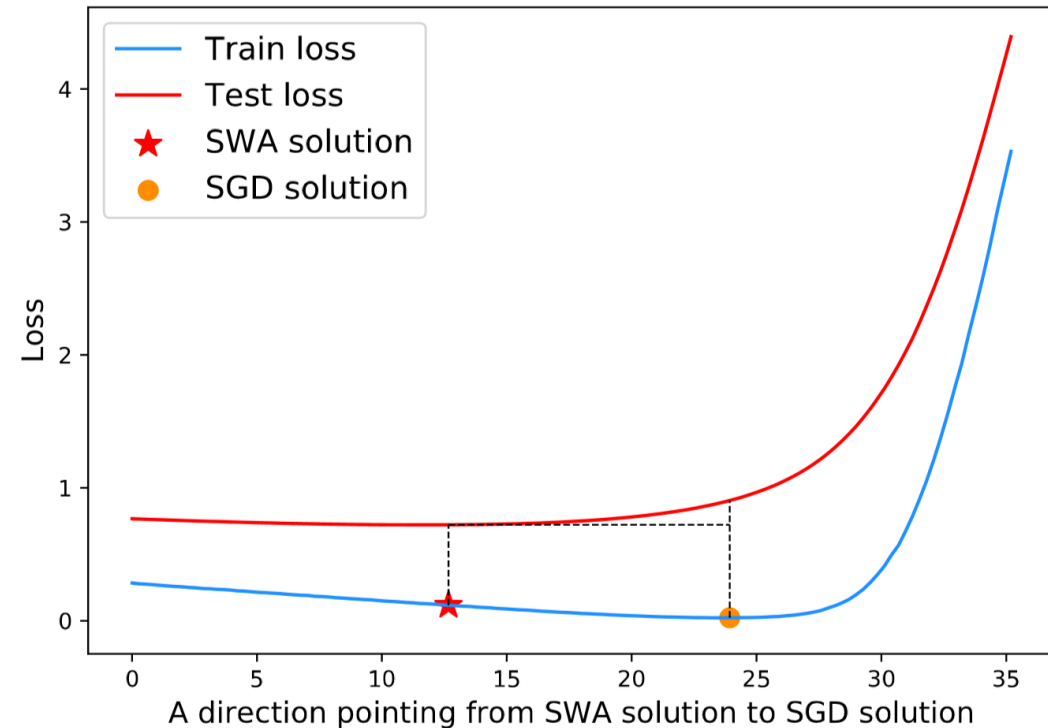


# Empirical Observation



Averaging SGD weight (SWA\*) indeed finds a biased solution with **higher training loss** but **lower test loss**.

This phenomenon can **NOT** be well explained by the “flatness/sharpness” theory!



\* Averaging weights leads to wider optima and better generalization. UAI Press, 2018.



## Leveraging asymmetric valleys (AVs):

- Designing new algorithms (e.g., SWA) based on our theory and intuition.
- Using the concept of AVs to explain which can not be explained by sharpness/flatness theory.

## Understanding asymmetric valleys (AVs):

- Where AVs originate from?
- What network structure or loss function tend to cause AVs





清华大学  
Tsinghua University

# Thanks

## Asymmetric Valleys: Beyond Sharp and Flat Local Minima

Haowei He | Gao Huang | Yang Yuan

Poster: Tue Dec 10th 05:30 -- 07:30 PM @ East Exhibition Hall B + C #116

December 10th, 2019