Target Contrastive Estimation

Wouter Kouw & Marco Loog
Pattern Recognition Lab

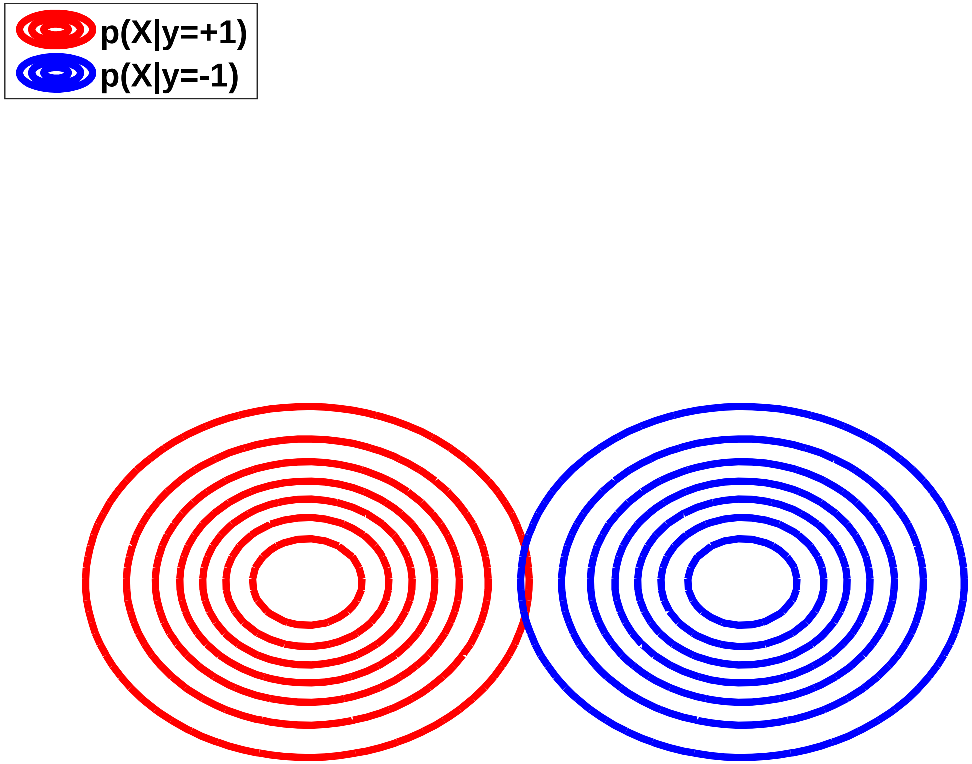
Domain Adaptation

- **A supervised learning setting where training and test data stem from different biased samplings.**

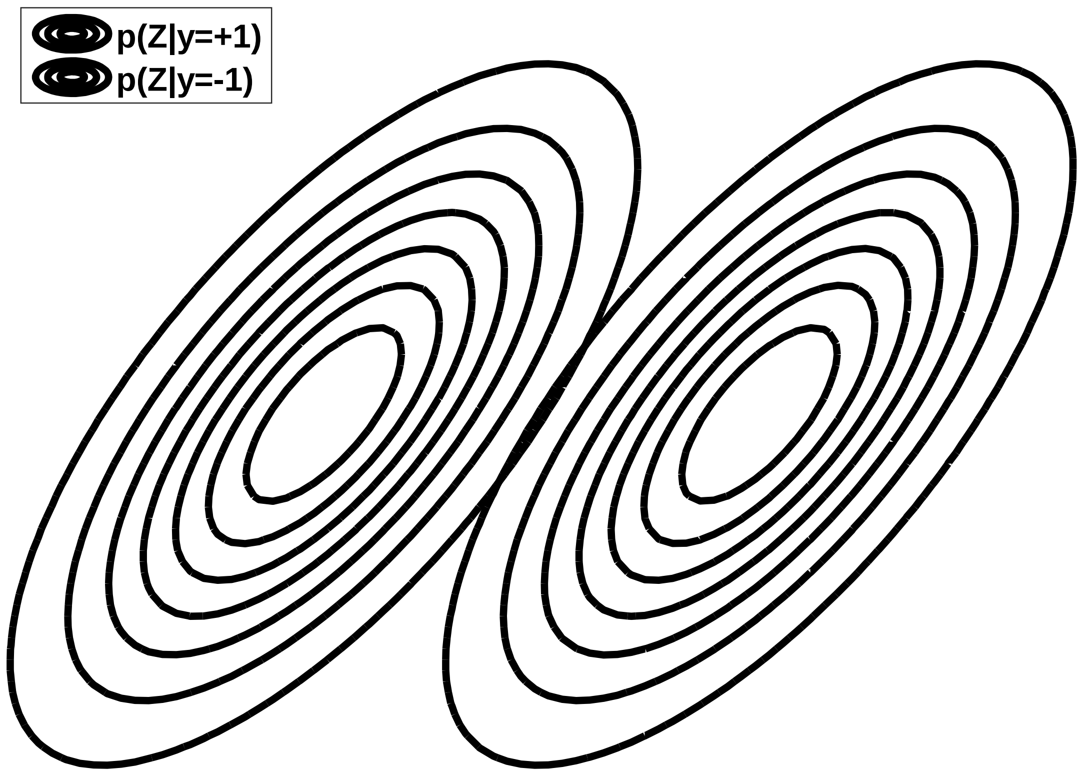
- **For example, perform the same clinical experiment in different hospitals.**
  - Geographically biased sampling.

- **More formally:**
  - Shared sample space Ω
  - Shared event space $\mathcal{F}$
  - Different probability measures $\mathbb{Q}$, $\mathbb{P}$

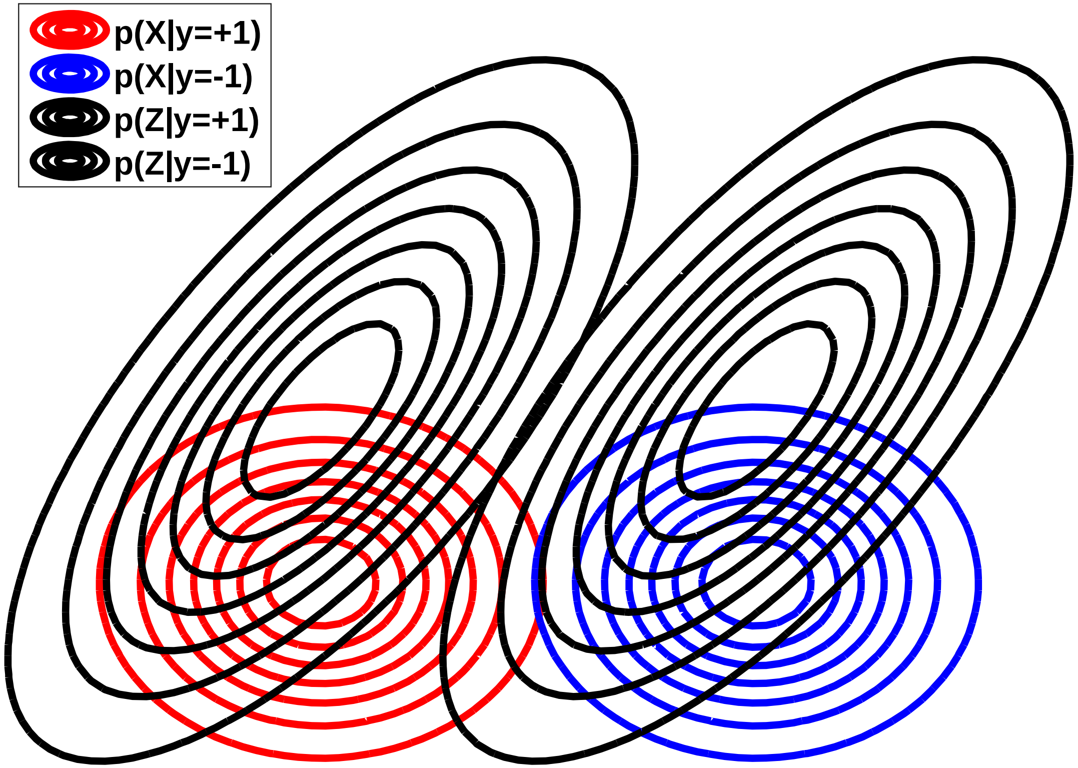
Domain Adaptation



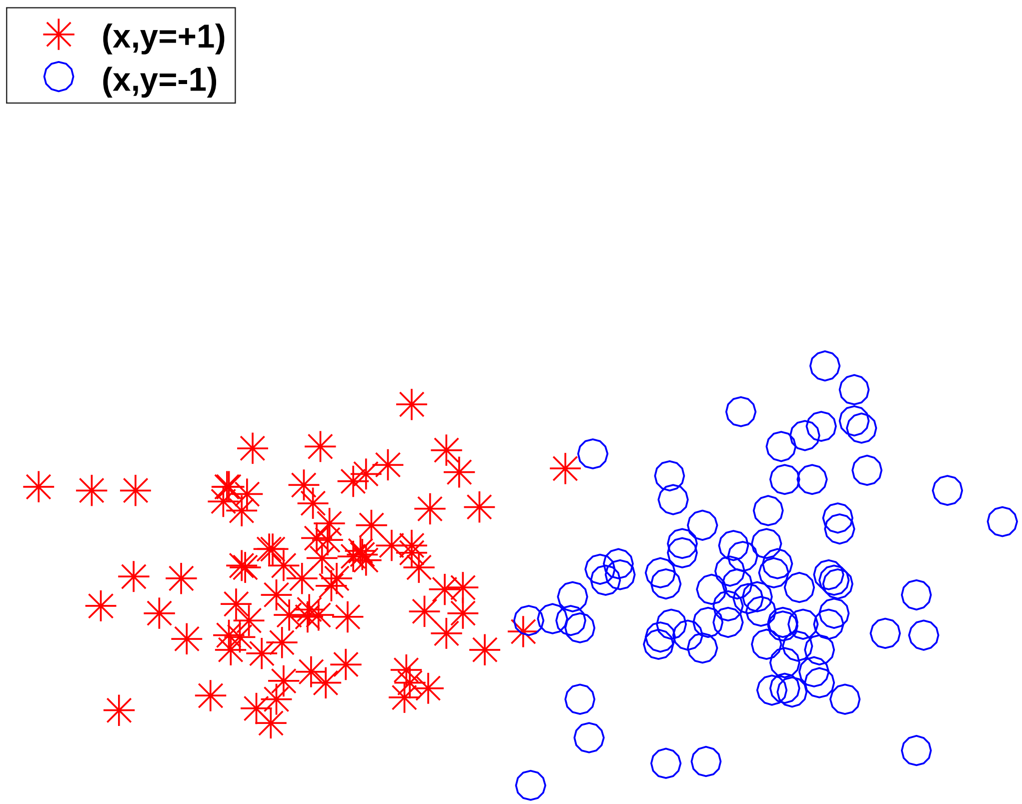
Domain Adaptation



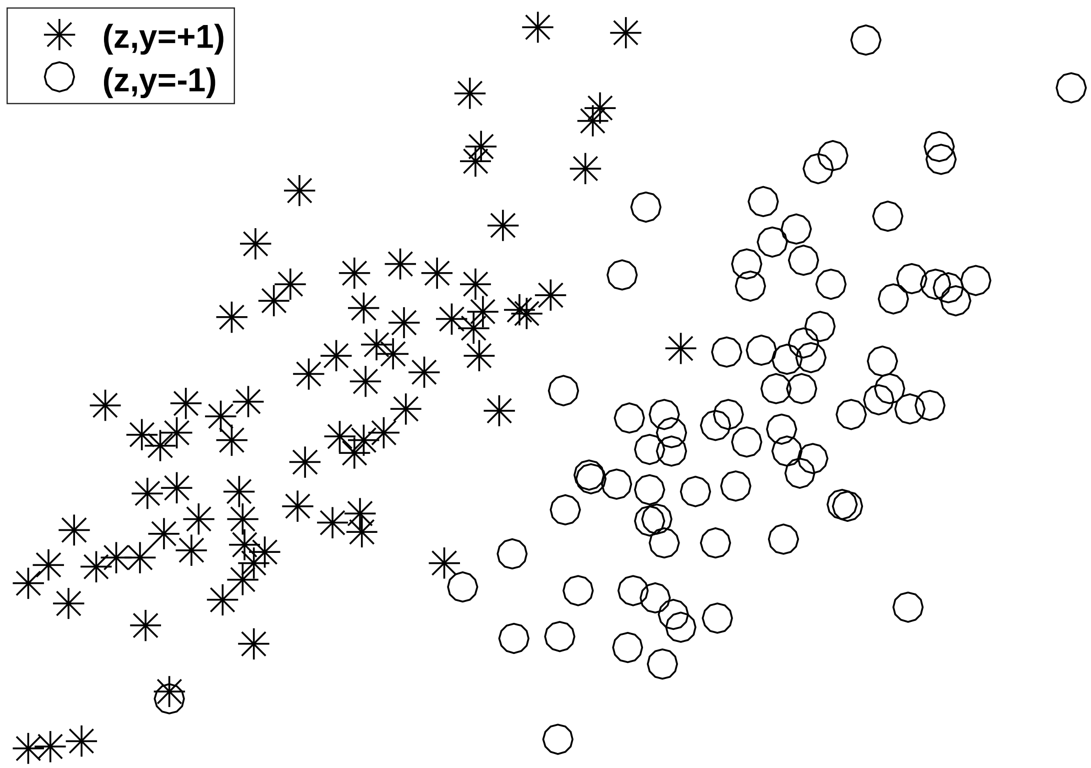
Domain Adaptation



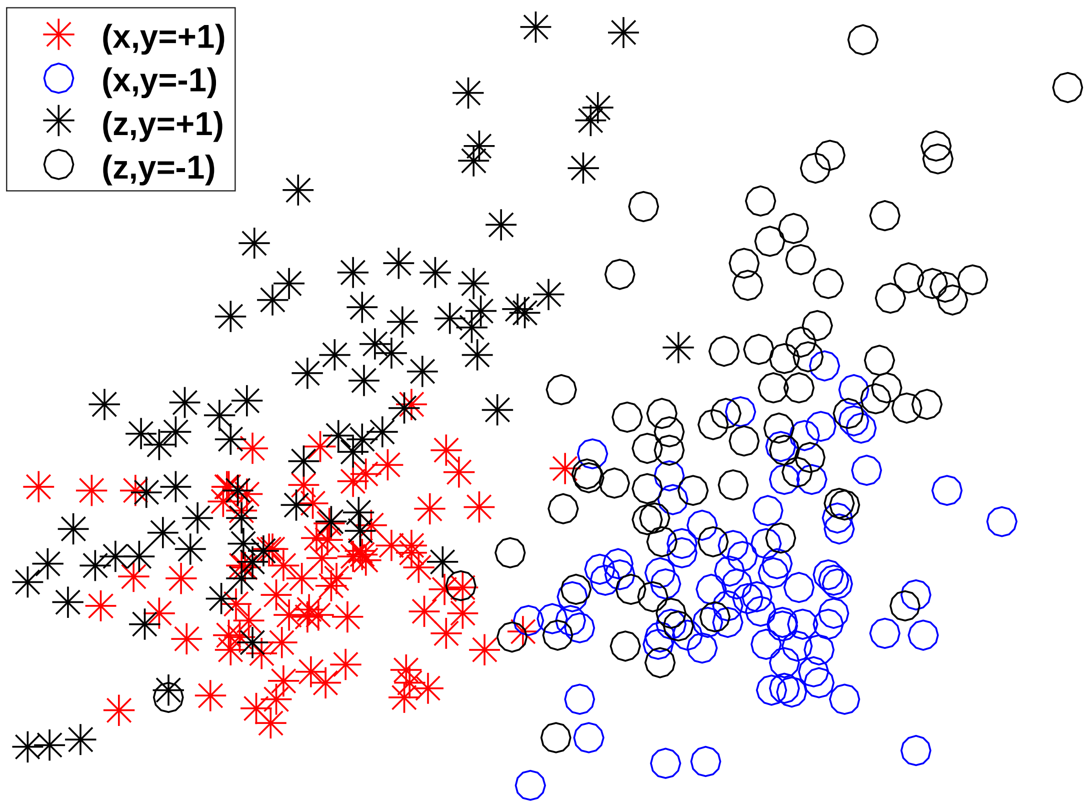
Domain Adaptation



Domain Adaptation



Domain Adaptation



Current approaches

- **A standard procedure for DA relies on making an assumption on domain dissimilarity and deviating from the source model.**

- **Sensitive to estimation errors.**
- **Sensitive to class-dependent transformations.**
- **Sensitive to disjoint empirical supports.**
- **Sensitive to model misspecification.**

- **As a result, DA approaches can perform worse than naive models.**

Target Contrastive Estimation

- **We are specifically interested in never performing worse than the source model.**

- **Can we construct a parameter estimator such that its likelihood is larger than or equal to the likelihood of the source estimator on the target domain?**

- **In order to do this, we will contrast the hypothetical target estimate with the source estimate for worst-case labellings.**

Notation

- **Source samples:**

$$x_i \in X = \{(x_{i1}, \dots, x_{id}) \in \mathbb{R}^d \mid x_i \sim p_X, i = 1, \dots, n\}$$

$$y_i \in y = \{y_i \in \{1, \dots, K\} \mid i = 1, \dots, n\}$$

- **Target samples:**

$$z_j \in Z = \{(z_{j1}, \dots, z_{jd}) \in \mathbb{R}^d \mid z_j \sim p_Z, j = 1, \dots, m\}$$

$$u_j \in u = \{u_j \in \{1, \dots, K\} \mid j = 1, \dots, m\}$$

- **Likelihood of model parameter given source data**

$$L(\theta \mid X, y)$$

- **Likelihood of model parameter given target data**

$$L(\theta \mid Z, u)$$

Source estimator

- **Since we have labeled source data, we can fit a model:**

$$\hat{\theta}_S = \underset{\theta \in \Theta}{\arg\max}\ L(\theta \mid X, y)$$

where Θ is the parameter space.

- **The likelihood of this parameter on the target samples can be evaluated through:**

$$L(\hat{\theta}_S \mid Z, u)$$

Contrast

- **We want to construct a parameter estimator $\hat{\theta}_T$ for which the following holds:**

$$L(\hat{\theta}_T \mid Z, u) \geq L(\hat{\theta}_S \mid Z, u)$$

or equivalently:

$$L(\hat{\theta}_T \mid Z, u) - L(\hat{\theta}_S \mid Z, u) \geq 0$$

- **Maximizing this contrast w.r.t. θ leads to an estimator that returns the source estimate when it can not do better.**

$$\max_{\theta \in \Theta} L(\theta \mid Z, u) - L(\hat{\theta}_S \mid Z, u) \geq 0$$

Pessimism

- **However, the true target labels u are unknown.**
- **In order to still construct an estimator that never performs worse than the source estimator, we can consider worst-case labellings.**
- **Such a labeling can be obtained by proposing a hypothetical labeling q_j for each sample and minimizing the likelihood:**

$$\min_q \, L(\theta \mid Z, q)$$

Pessimism

- **Incorporating the minimization over labellings in the contrast yields:**

$$\max_{\theta \in \Theta} \min_{q} L(\theta \mid Z, q) - L(\hat{\theta}_S \mid Z, q) \geq 0$$

- **If we choose discrete labellings, the minimization will be combinatorial and expensive.**
  - Therefore, we employ a convex relaxation of the labeling space.
  - Corresponds to class posterior probabilities: $q_{kj} := p(u_j = k \mid z_j)$.
  - q_j will be an element of a K-1 dimensional simplex Δ_{K-1} .

Target Contrastive Estimation

– **The resulting maximum contrastive pessimistic likelihood estimator is :**

$$\hat{\theta}_T = \underset{\theta \in \Theta}{\arg\max} \min_{q \in \Delta^m_{K-1}} L(\theta \mid Z, q) - L(\hat{\theta}_S \mid Z, q)$$

Linear Discriminant

- **Linear Discriminant Analysis is a classical classifier with a particularly interesting property under this estimator.**

- **LDA fits a Gaussian distribution to each class:**

$$L(\theta \mid Z, q) = \sum_{j=1}^{m} \sum_{k=1}^{K} q_{kj} \log \pi_k \mathcal{N}(z_j \mid \mu_k, \Sigma)$$

where $\theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma)$

- **Plugging that into the TCE formulation:**

$$\hat{\theta}_T = \underset{\theta \in \Theta}{\arg\max} \min_{q \in \Delta_{K-1}^{m}} \sum_{j=1}^{m} \sum_{k=1}^{K} q_{kj} \log \frac{\pi_k \mathcal{N}(z_j \mid \mu_k, \Sigma)}{\hat{\pi}_{Sk} \mathcal{N}(z_j \mid \hat{\mu}_{Sk}, \hat{\Sigma}_S)}$$

TCE-LDA

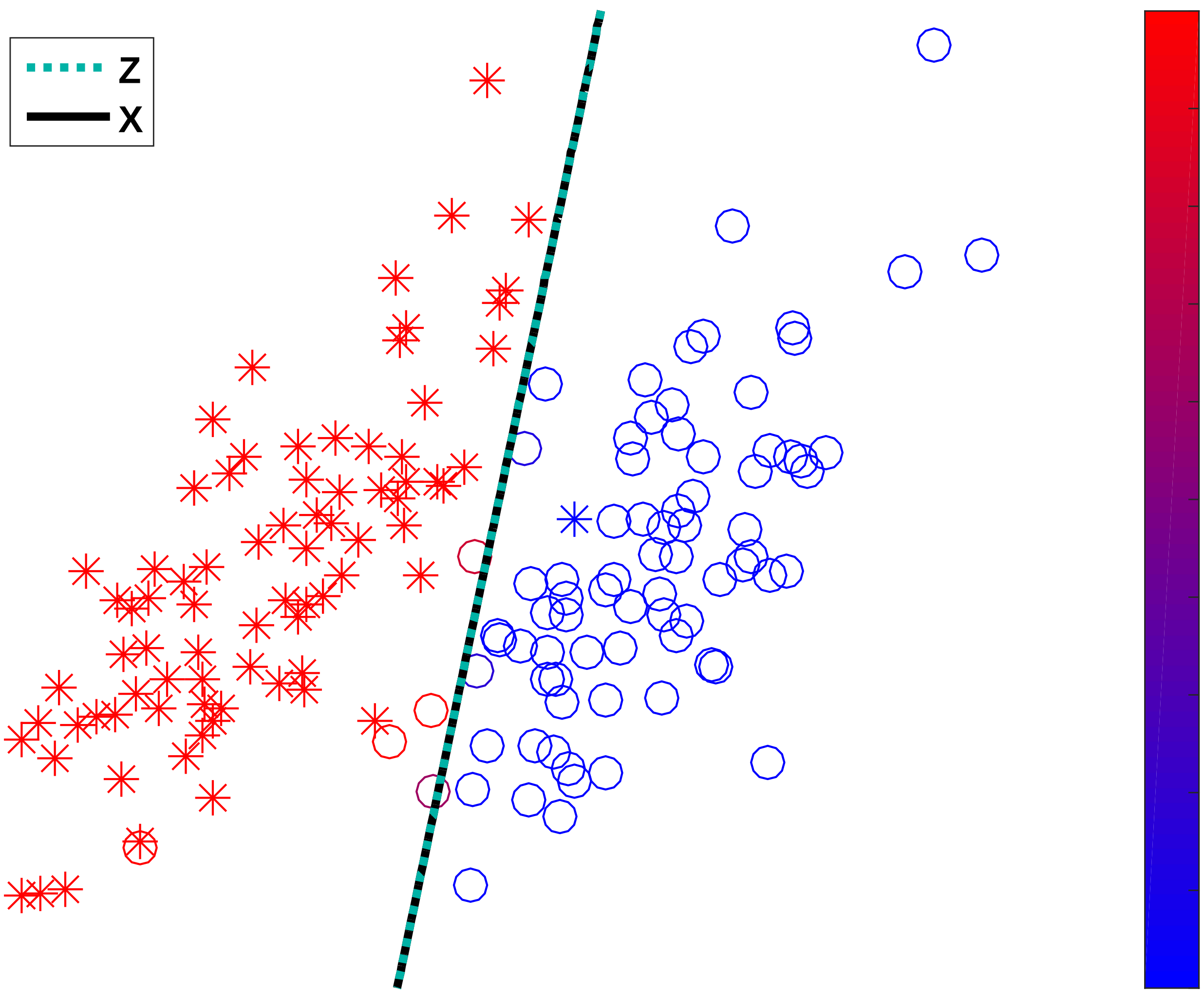
Theorem:

For continuously distributed feature vectors and a sample size $m \geq d + K$, the likelihood of the TCE estimate is almost surely strictly larger than the source estimate.

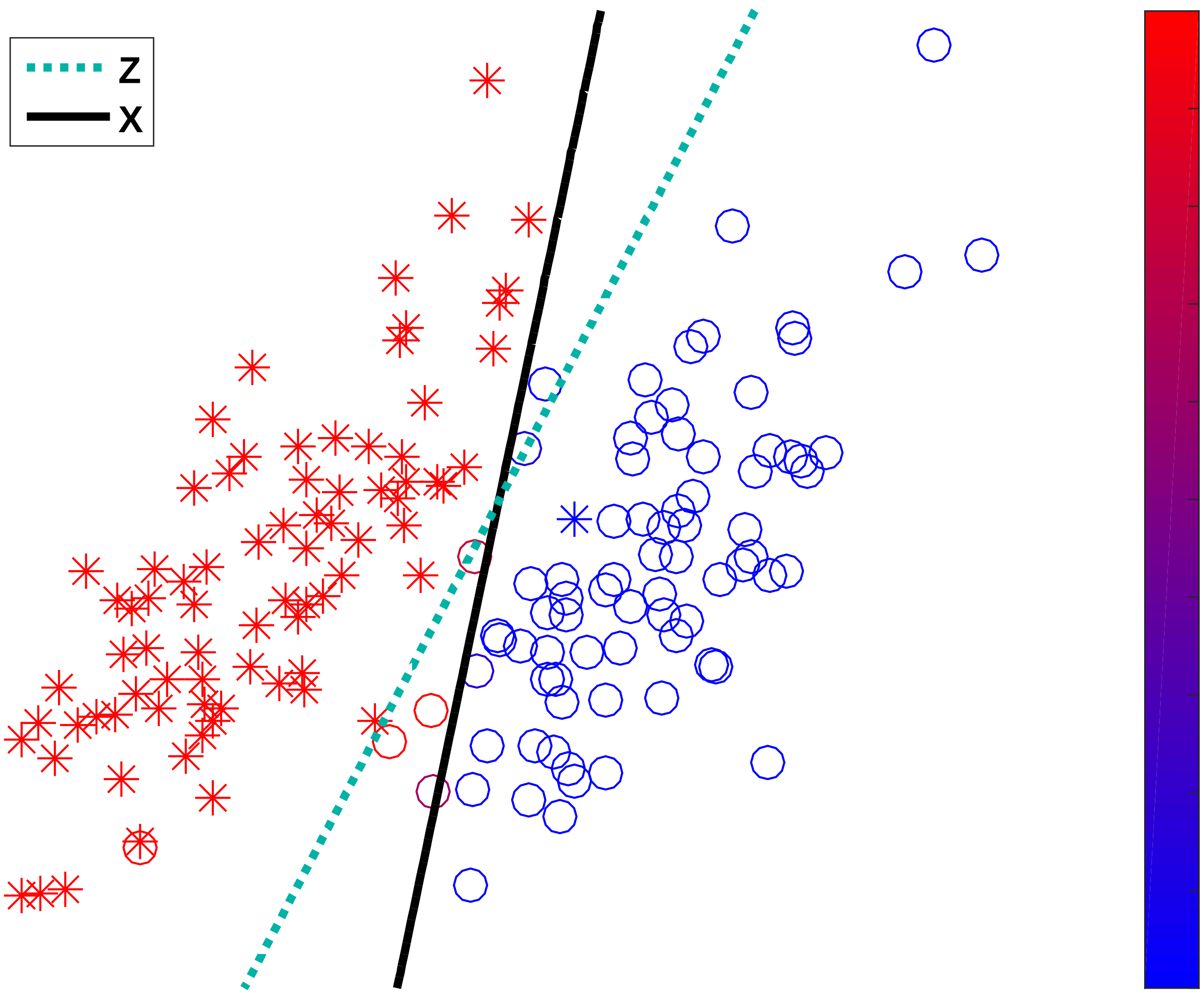
$$L_{\mathrm{LDA}}(\hat{\theta}_T \mid Z, u) > L_{\mathrm{LDA}}(\hat{\theta}_S \mid Z, u)$$

- **Does not hold for novel target samples.**
- **Does not directly translate to other measures (e.g. error rate).**

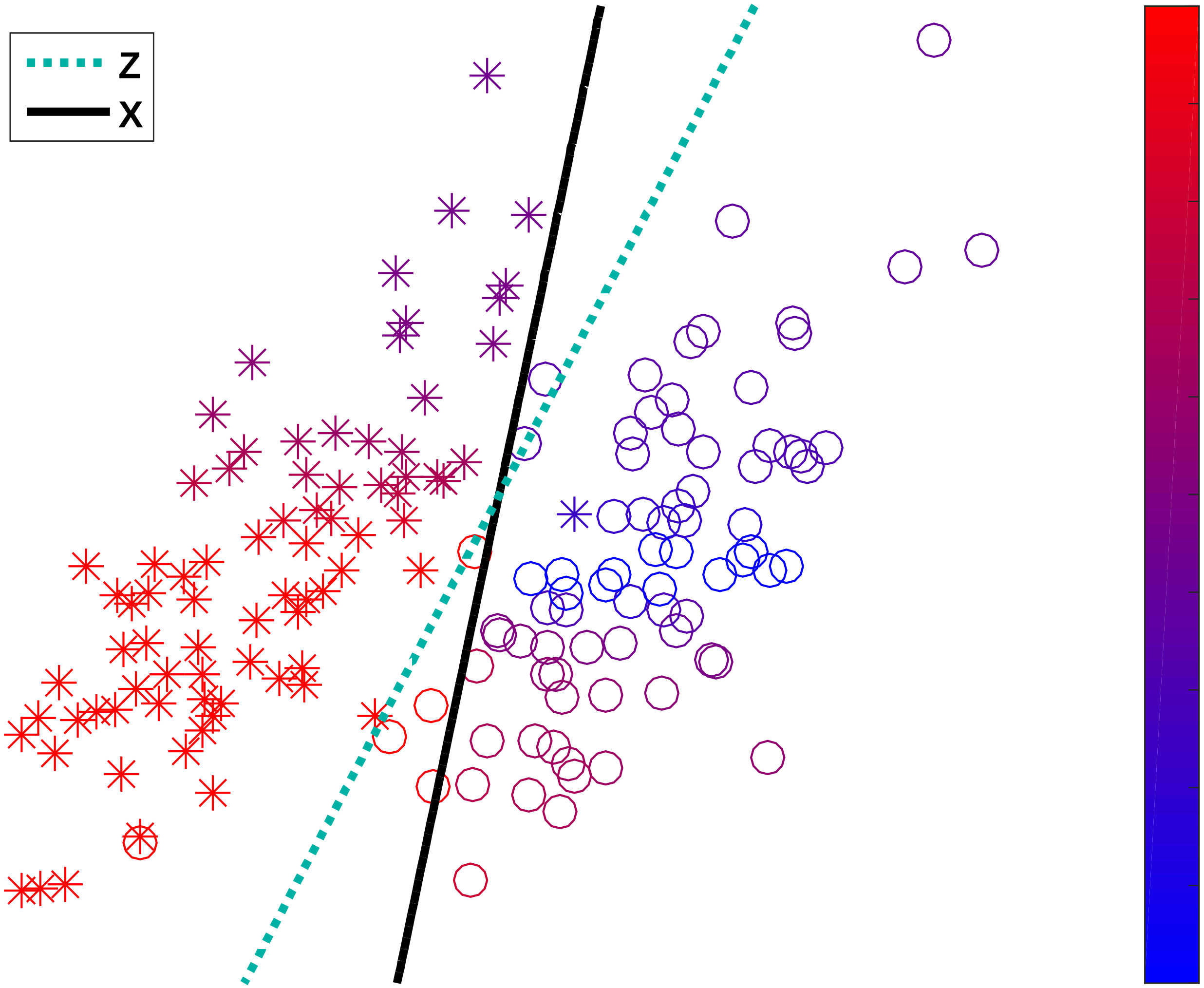
Optimization - init



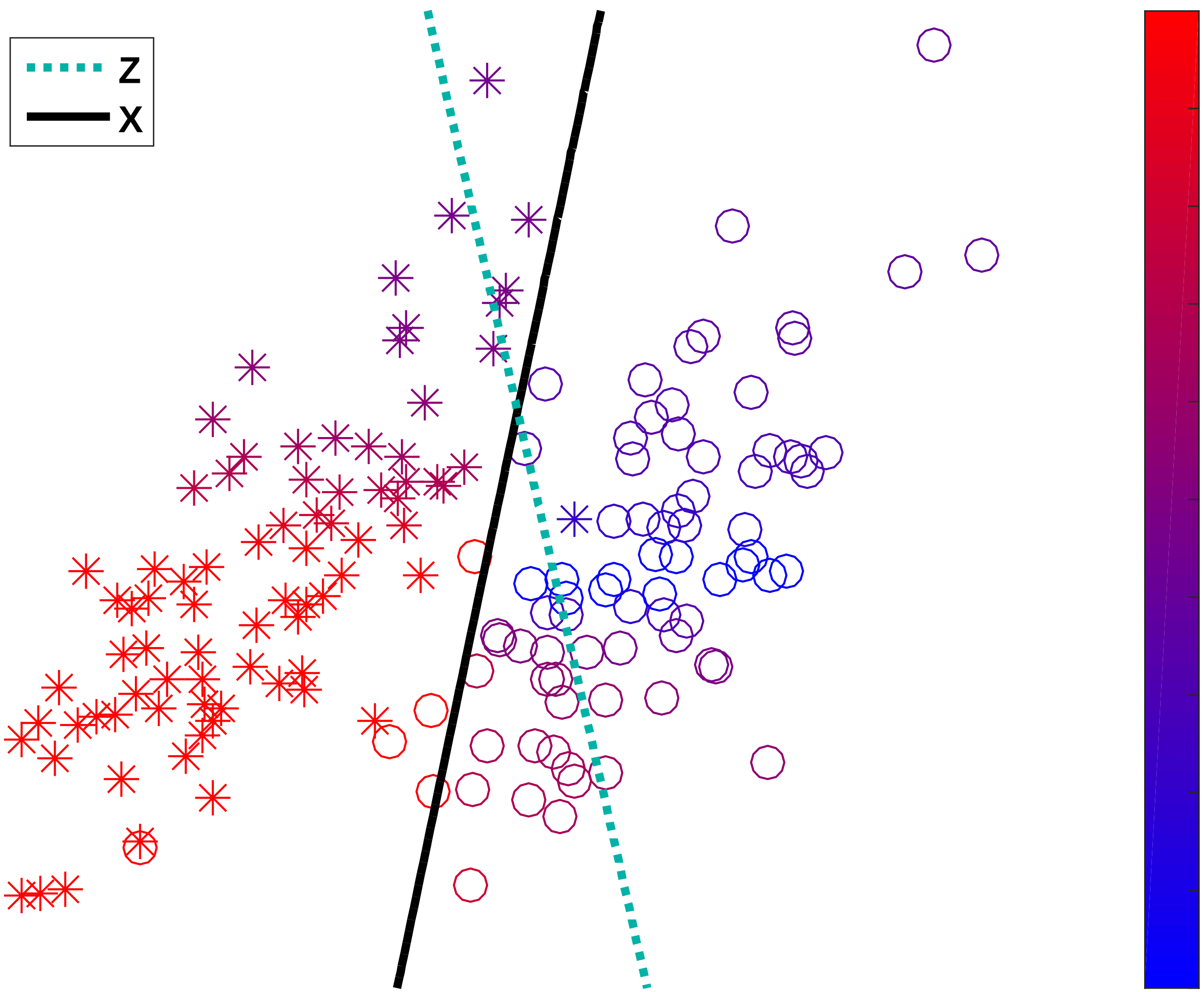
Optimization - max



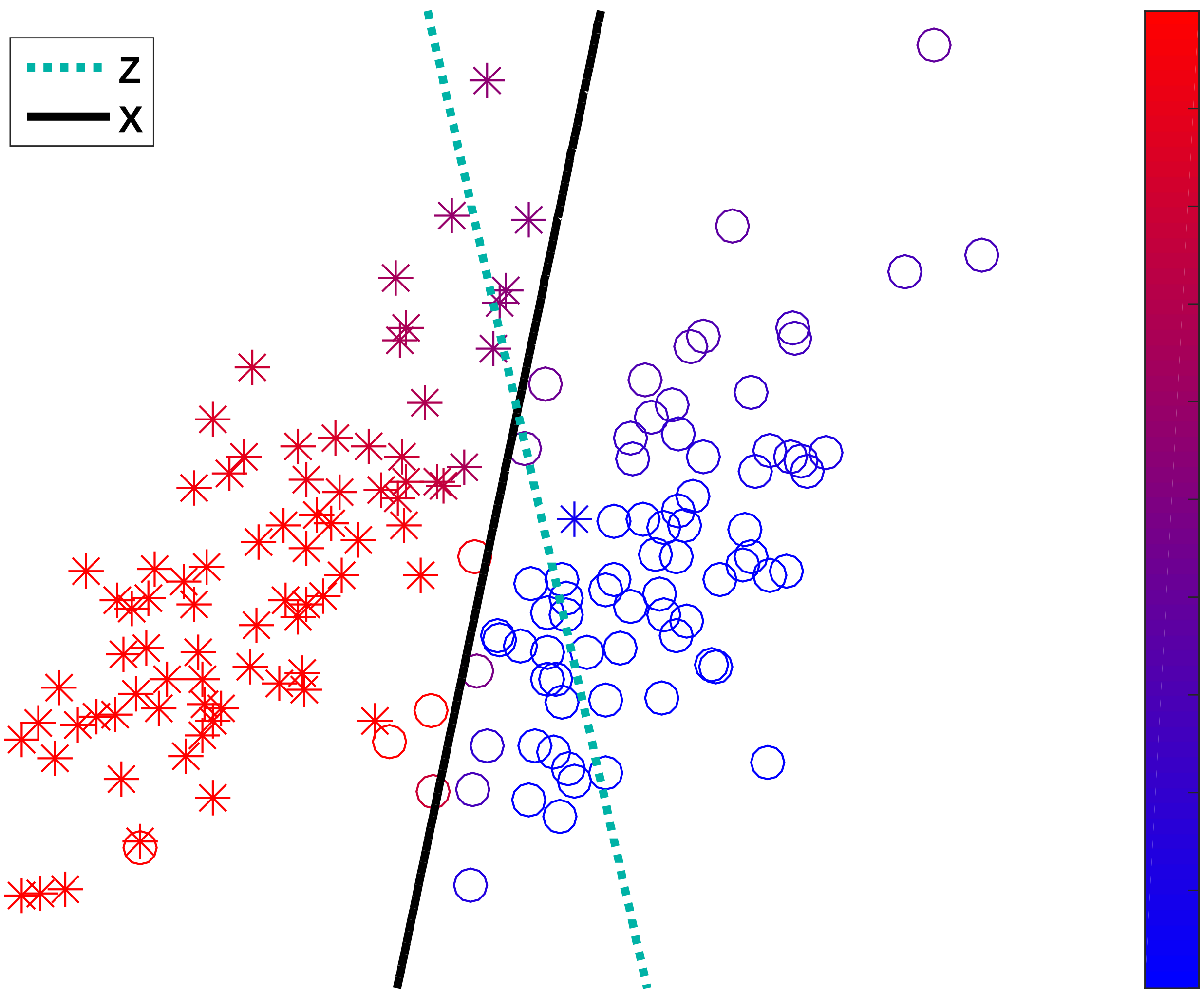
Optimization - min



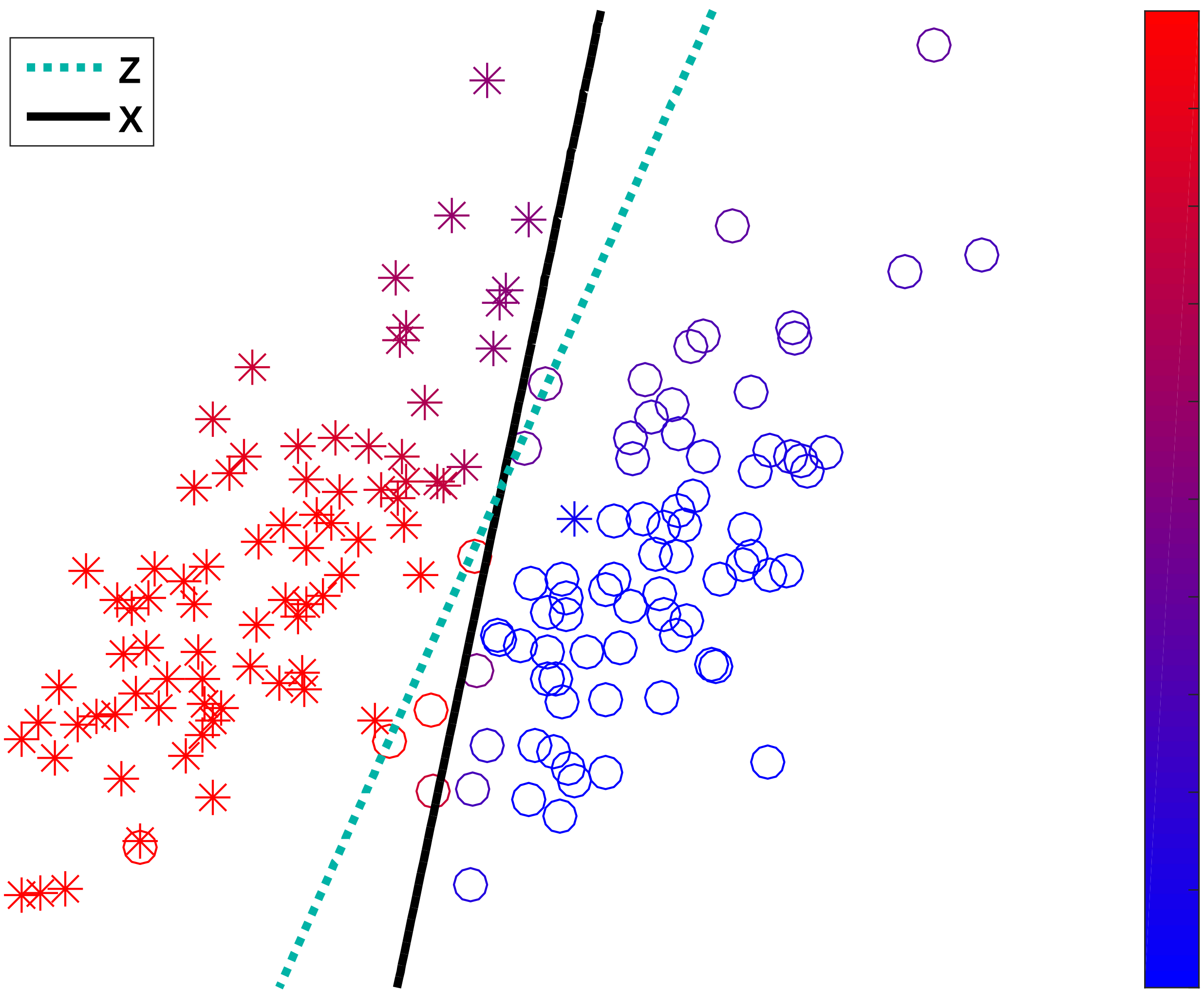
Optimization - max



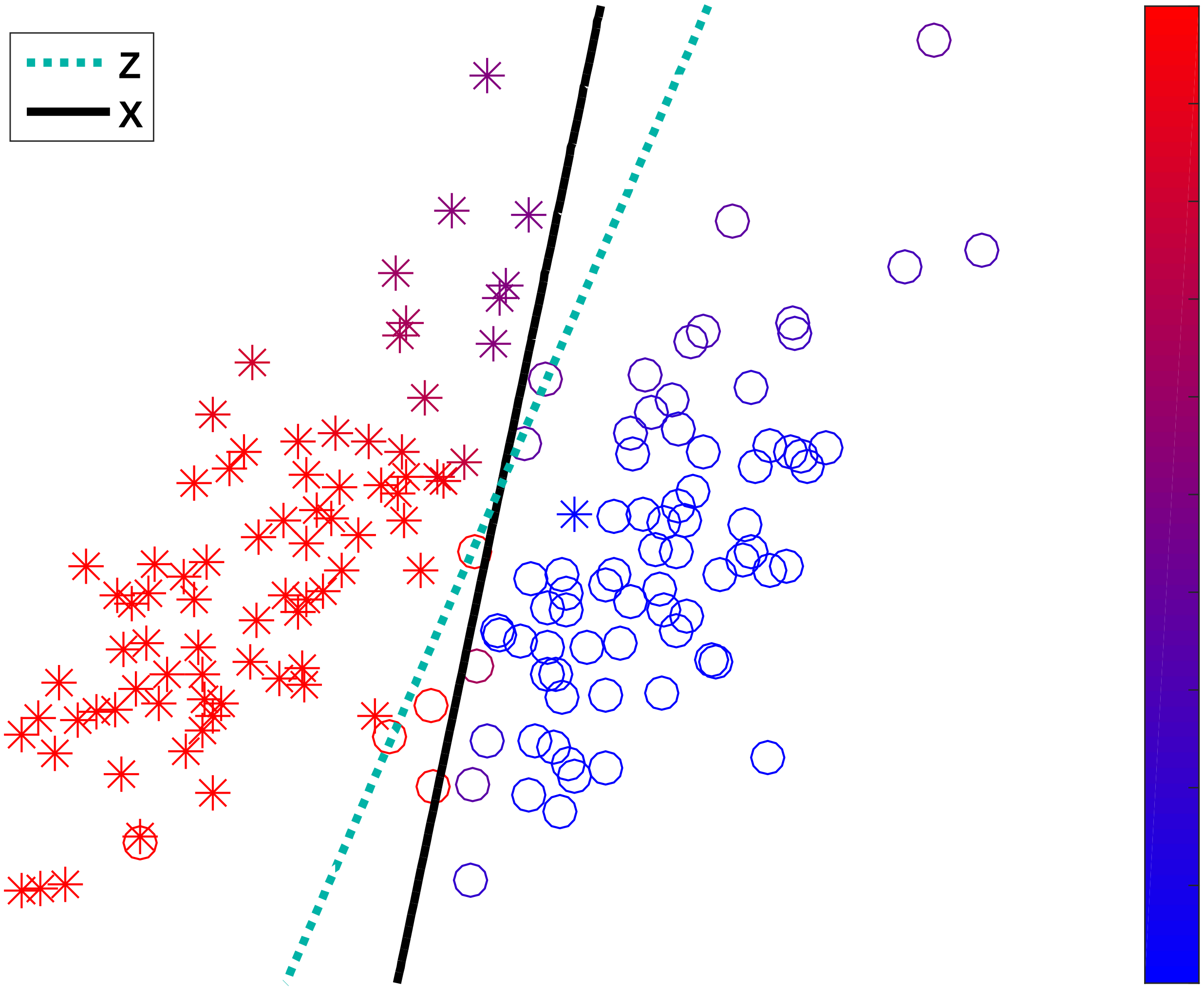
Optimization - min



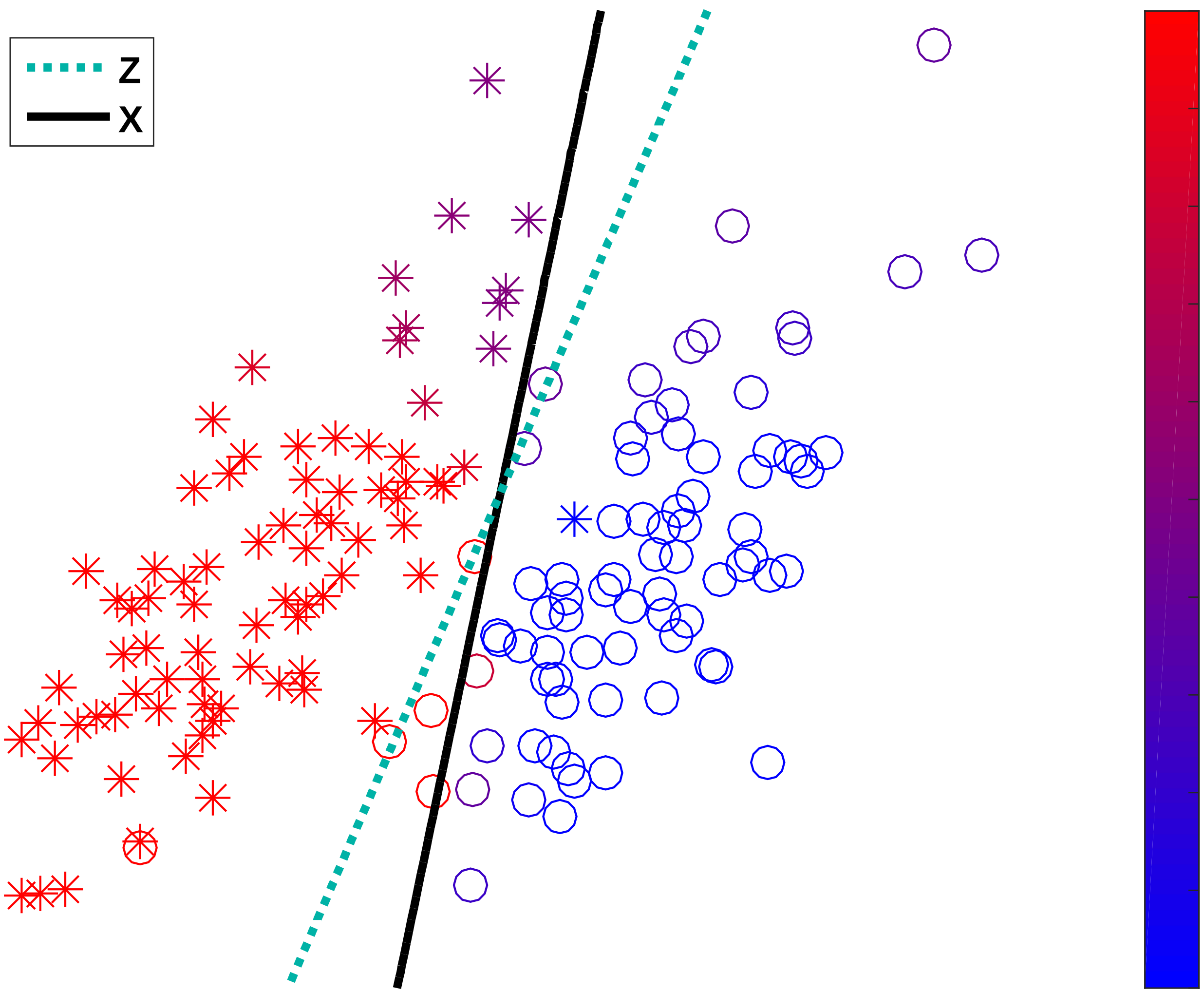
Optimization - max



Optimization - min



Optimization – saddle



Experiment

- **Heart disease dataset:**
  - 4 hospitals: Cleveland, Virginia, Hungary and Switzerland.

X	Z	S-LDA	TCE-LDA
C	V	$-3.68e+3$	$-3.19e+3$
C	H	$-4.97e+3$	$-4.69e+3$
C	S	$-2.20e+3$	$2.18e+2$
V	H	$-5.94e+3$	$-4.80e+3$
V	S	$-2.51e+3$	$3.47e+2$
H	S	$-2.45e+3$	$1.93e+2$

X	Z	S-LDA	TCE-LDA
V	C	$-6.41e+3$	$-5.08e+3$
H	C	$-5.35e+3$	$-5.00e+3$
S	C	$-3.13e+18$	$-5.16e+3$
H	V	$-3.60e+3$	$-3.25e+3$
S	V	$-2.06e+18$	$-3.14e+3$
S	H	$-3.04e+18$	$-4.94e+3$

Experiment

- **Heart disease dataset:**
  - 4 hospitals: Cleveland, Virginia, Hungary and Switzerland.

X	Z	S-LDA	TCE-LDA
C	V	.410 (.035)	.395 (.035)
C	H	.174 (.022)	.174 (.022)
C	S	.463 (.045)	.455 (.045)
V	H	.221 (.024)	.231 (.025)
V	S	.366 (.043)	.366 (.045)
H	S	.545 (.045)	.537 (.045)

X	Z	S-LDA	TCE-LDA
V	C	.287 (.026)	.300 (.026)
H	C	.201 (.023)	.208 (.023)
S	C	.380 (.028)	.317 (.027)
H	V	.415 (.035)	.415 (.035)
S	V	.340 (.034)	.340 (.034)
S	H	.384 (.028)	.378 (.028)

Conclusion

- **Target Contrastive Estimation obtains parameters that are never less likely than the source estimators.**

- **Increases in likelihood do not directly correspond to decreases in error rates.**

- **The results due to the worst-case labeling do not hold for novel target samples (transductive only).**

Questions