Lower Bounds on Adversarial Robustness from Optimal Transport

Arjun Nitin Bhagoji[1]

Joint work with Daniel Cullina[2] and Prateek Mittal[1]

[1]Princeton University [2]Penn State

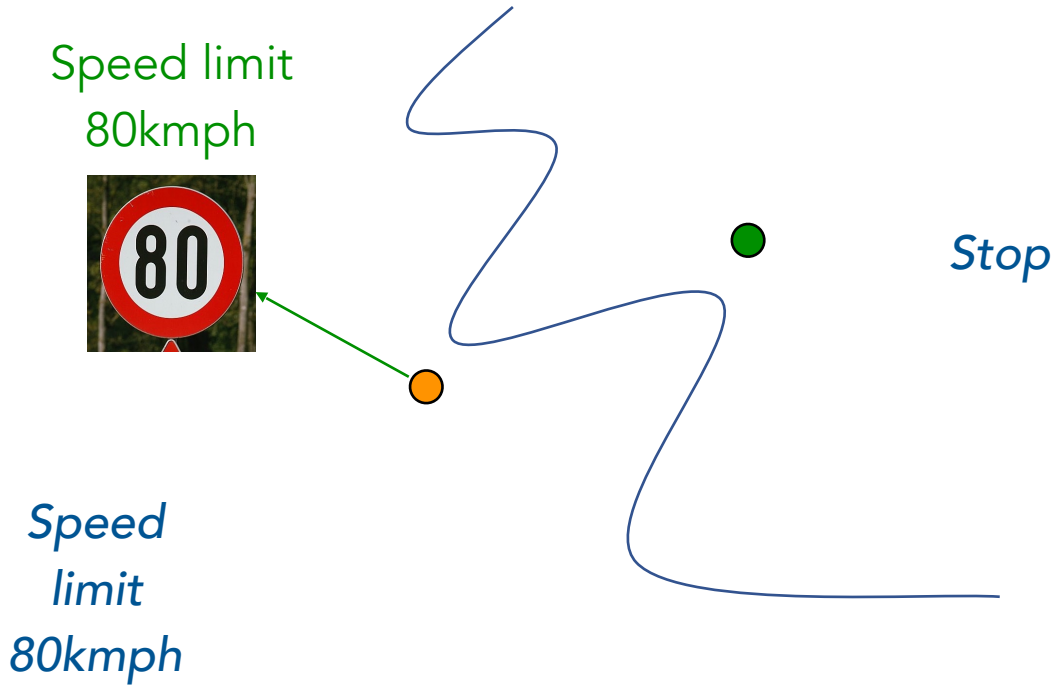
Reasoning about adversaries

Reasoning about adversaries

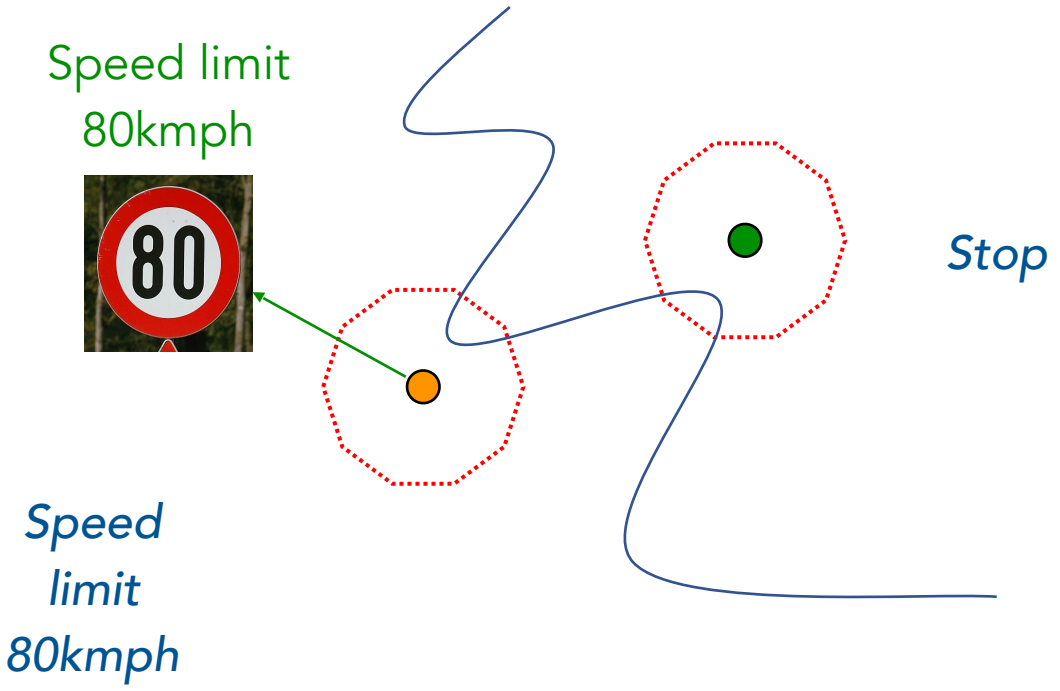
Speed limit 80kmph

Stop

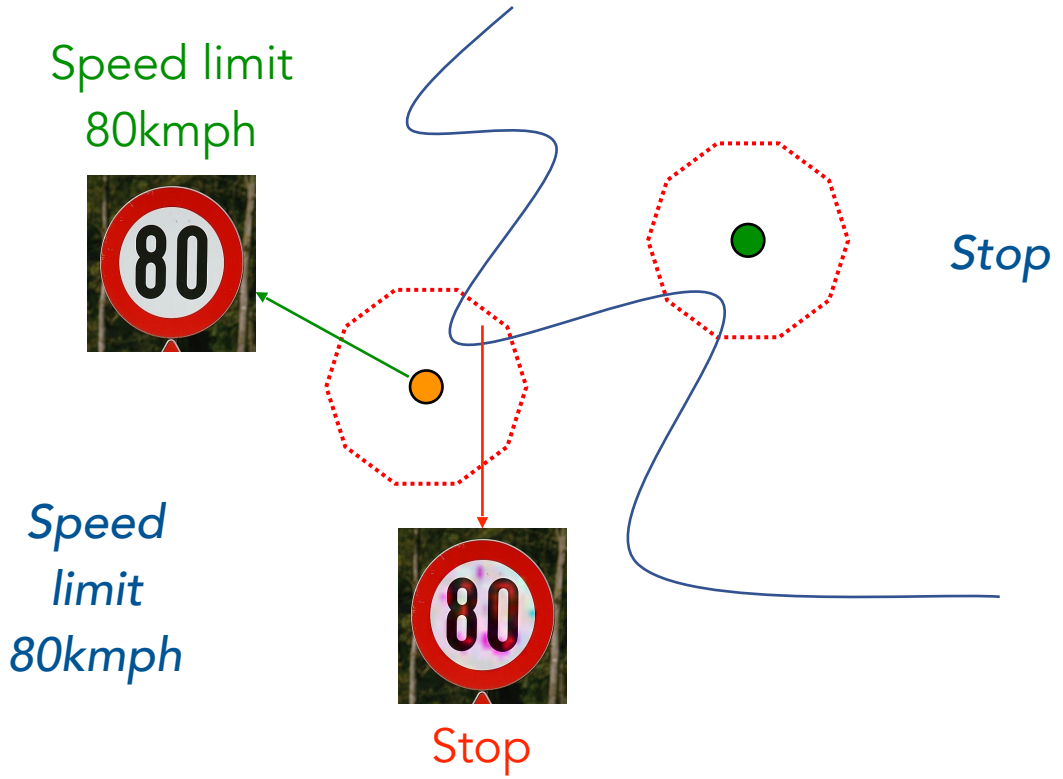
Reasoning about adversaries



Reasoning about adversaries

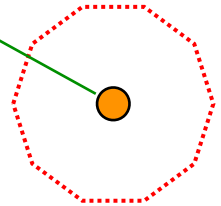


Reasoning about adversaries

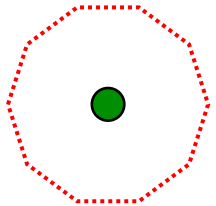


Reasoning about adversaries

Speed limit
80kmph

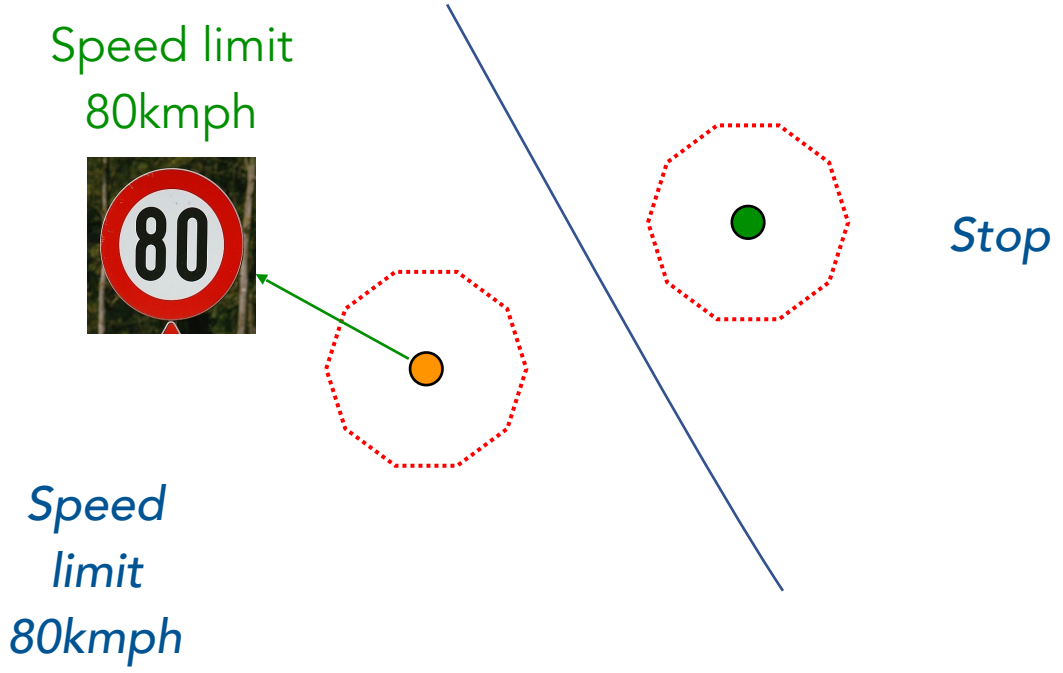


Speed limit 80kmph



Stop

Reasoning about adversaries



Can we argue rigorously about when a classifier exists that can distinguish between adversarially modified points?

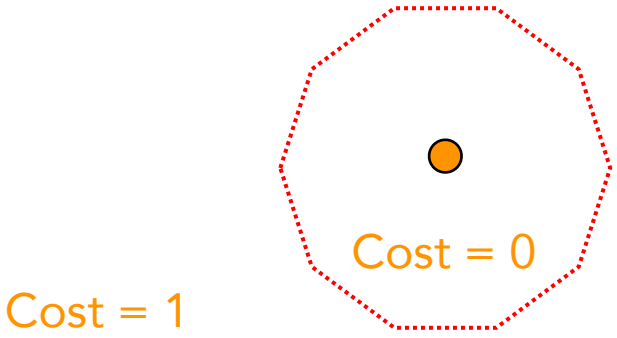
Intuition: Use data geometry

Intuition: Use data geometry

Space of data

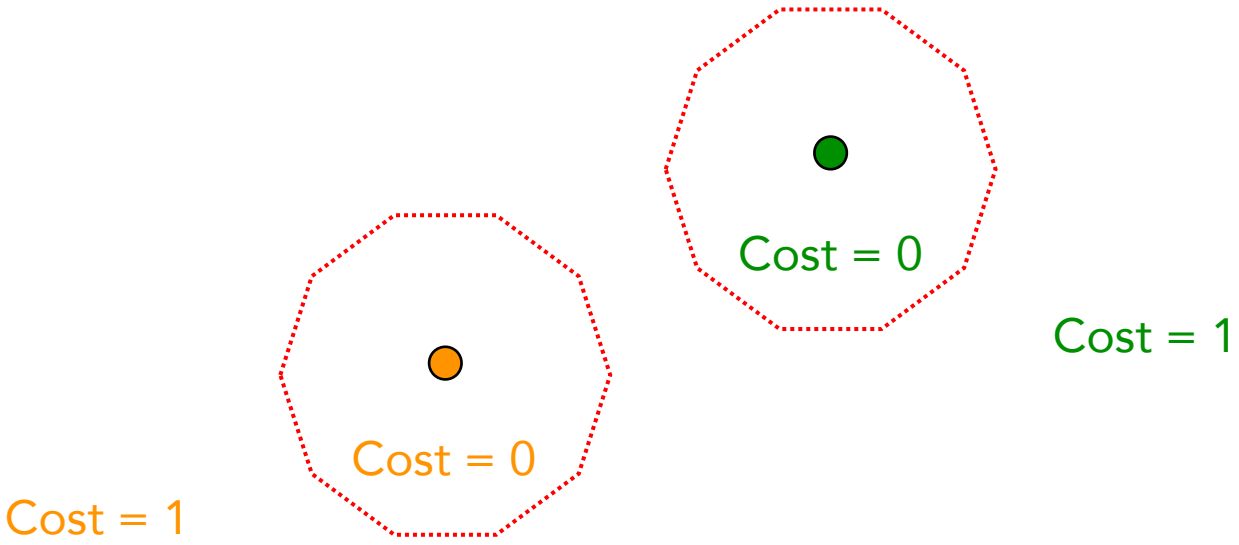
Intuition: Use data geometry

Space of data



Intuition: Use data geometry

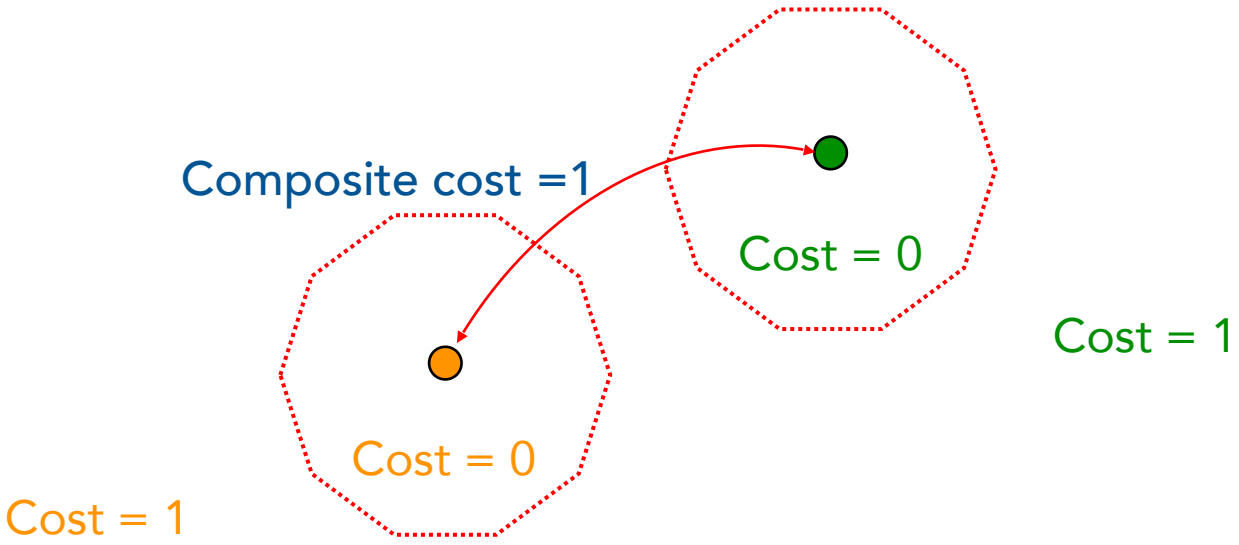
Space of data



Intuition: Use data geometry

Space of data

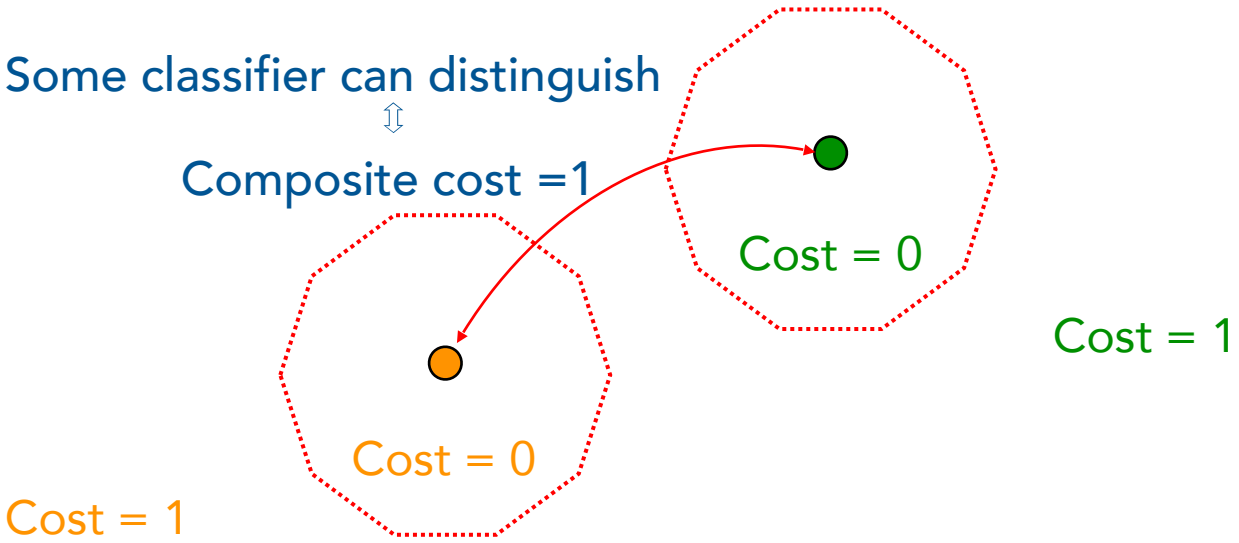
Note: Composite cost 0 only if cost 0 regions for two points intersect



Intuition: Use data geometry

Space of data

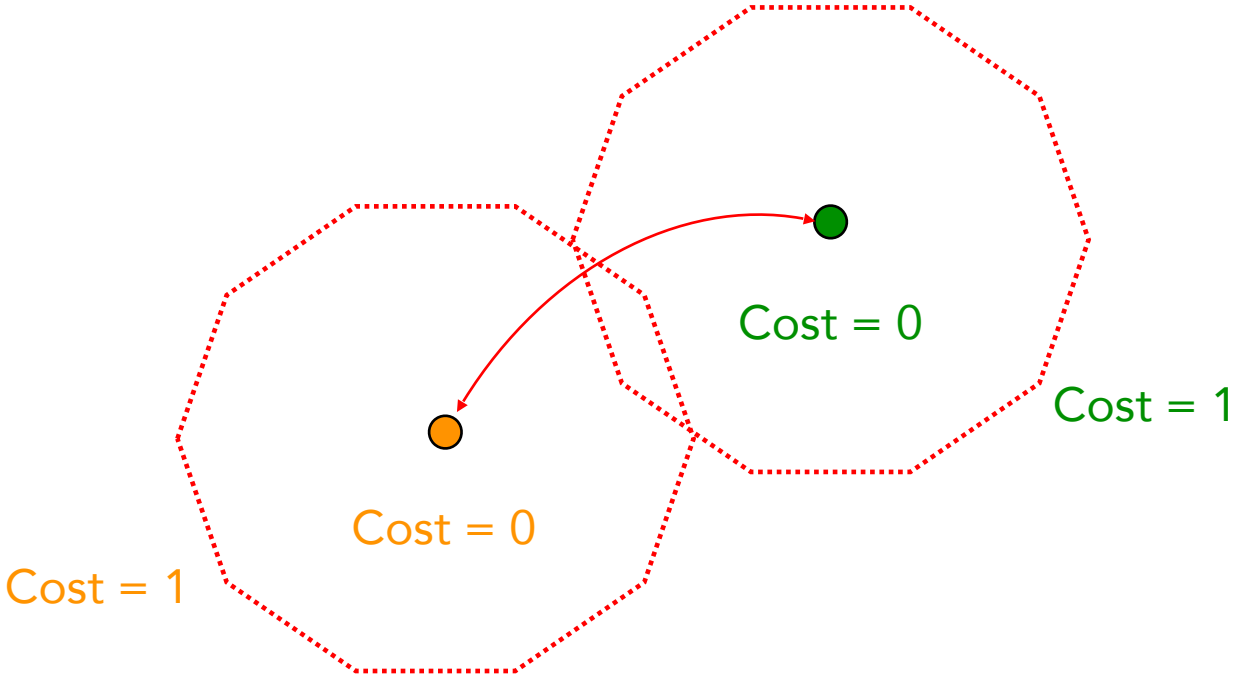
Note: Composite cost 0 only if cost 0 regions for two points intersect



Intuition: Use data geometry

Space of data

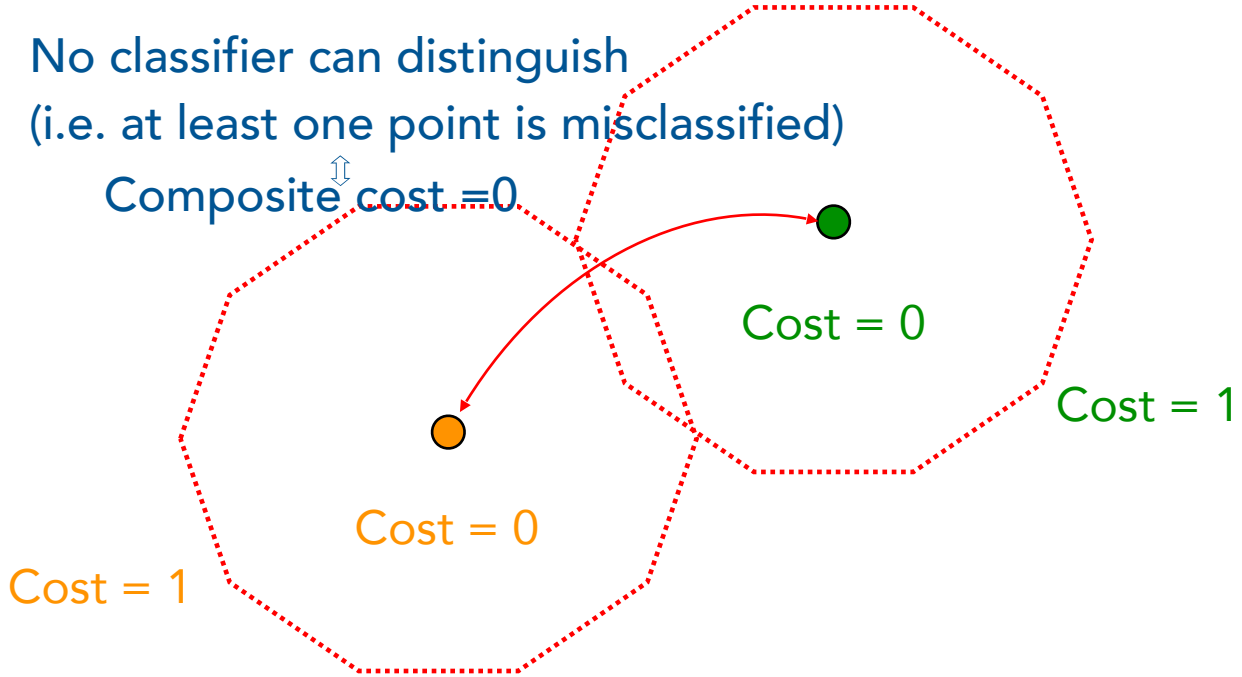
Note: Composite cost 0 only if cost 0 regions for two points intersect



Intuition: Use data geometry

Space of data

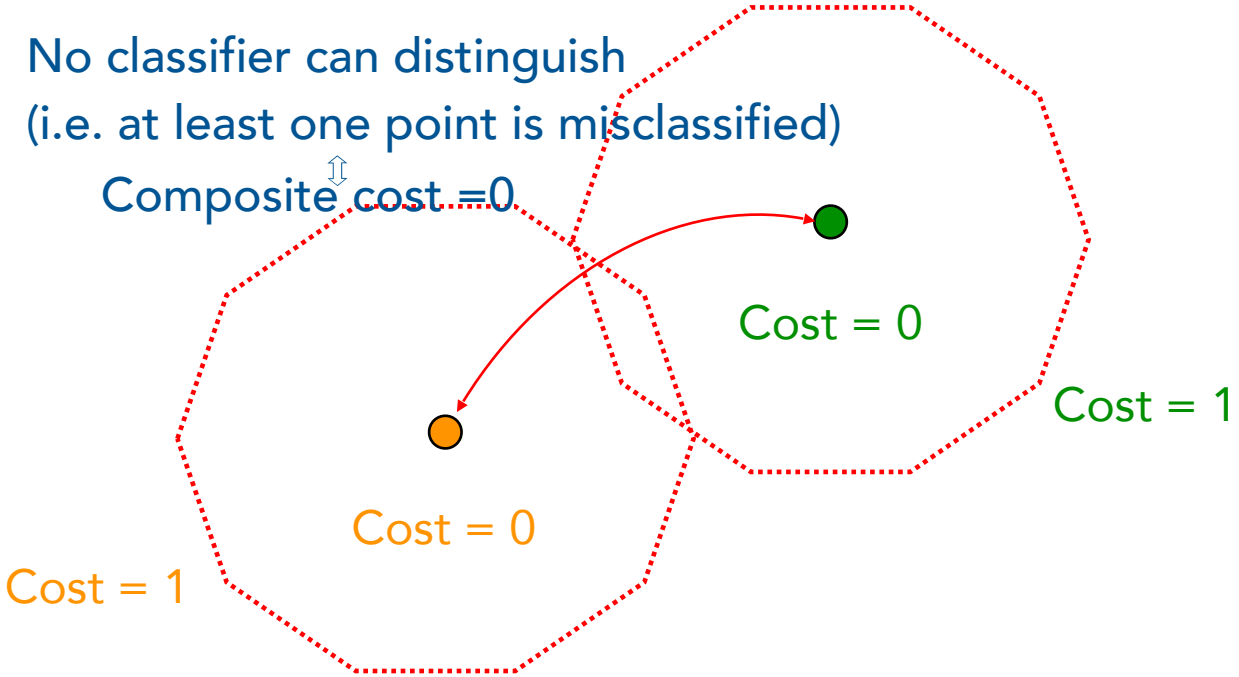
Note: Composite cost 0 only if cost 0 regions for two points intersect



Intuition: Use data geometry

Space of data

Note: Composite cost 0 only if cost 0 regions for two points intersect

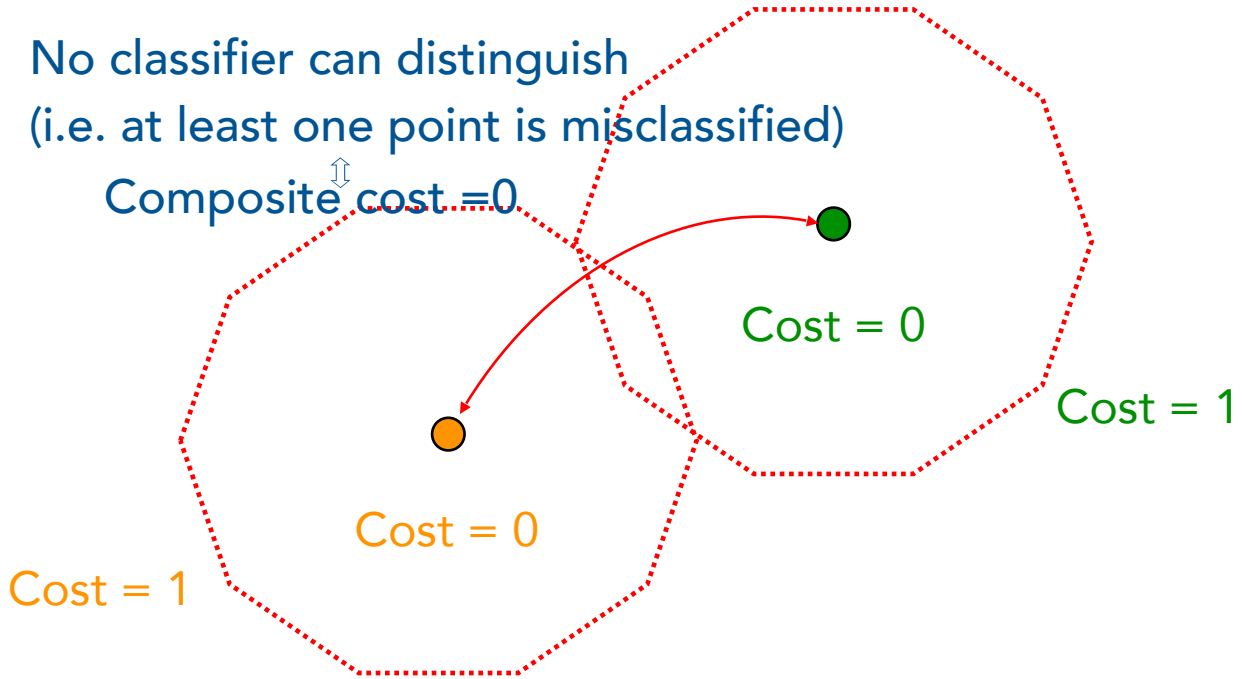


How do we compute the cost between class-wise distributions?

Intuition: Use data geometry

Space of data

Note: Composite cost 0 only if cost 0 regions for two points intersect



How do we compute the cost between class-wise distributions?

Find the joint distribution with the lowest cost: $C(P_1, P_{-1}) = \inf_{P_{1,-1} \in \Pi(P_1, P_{-1})} \mathbb{E}_{(X_1, X_{-1}) \sim P_{1,-1}}[c(X_1, X_{-1})]$

Discrete distributions: Minimum weight matching

Discrete distributions: Minimum weight matching

Consider the following i.i.d distributions on the line

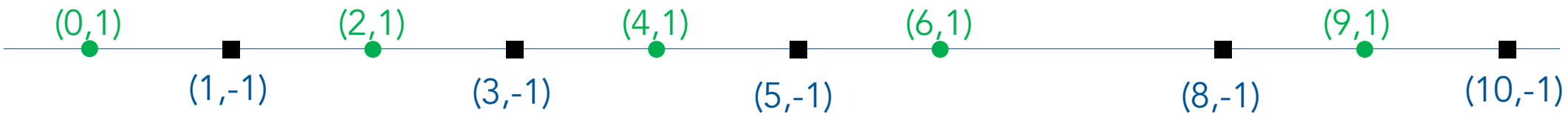
Discrete distributions: Minimum weight matching

Consider the following i.i.d distributions on the line

Discrete distributions: Minimum weight matching

Consider the following i.i.d distributions on the line

Without an adversary, optimal classifier loss is 0

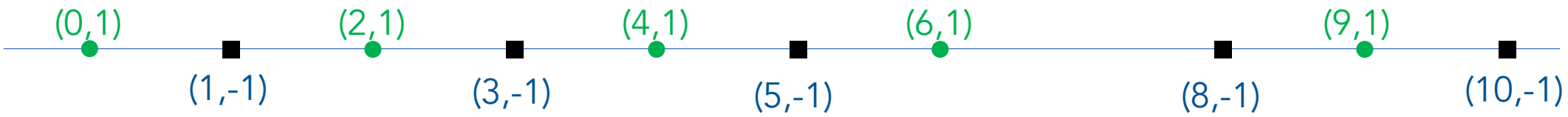


Discrete distributions: Minimum weight matching

Consider the following i.i.d distributions on the line

Without an adversary, optimal classifier loss is 0

Let the adversary have a budget $\beta = 0.55$

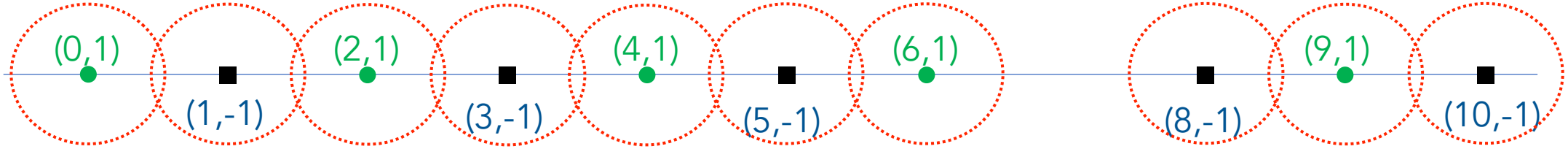


Discrete distributions: Minimum weight matching

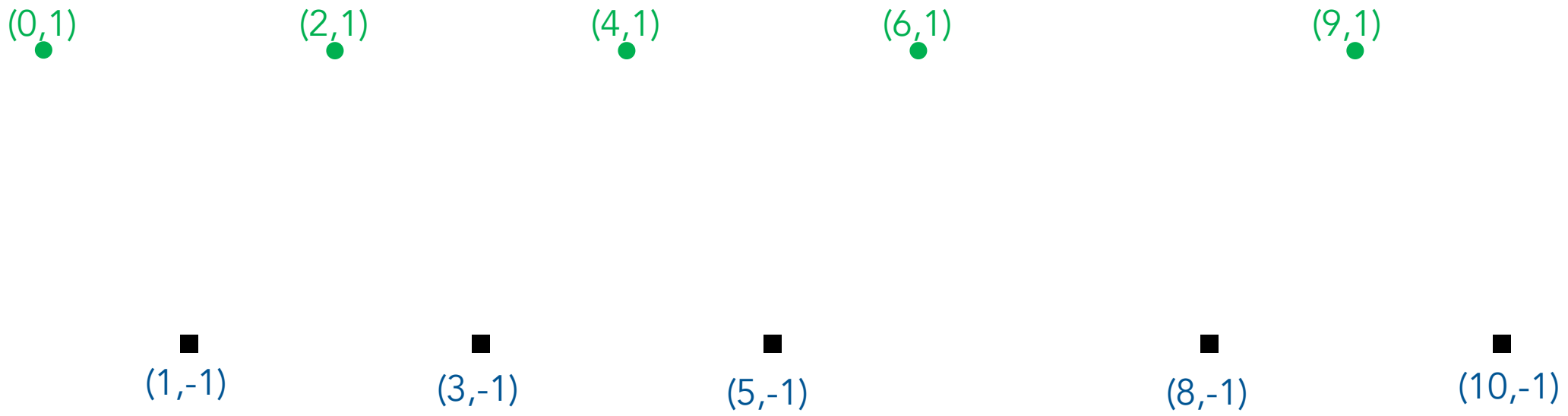
Consider the following i.i.d distributions on the line

Without an adversary, optimal classifier loss is 0

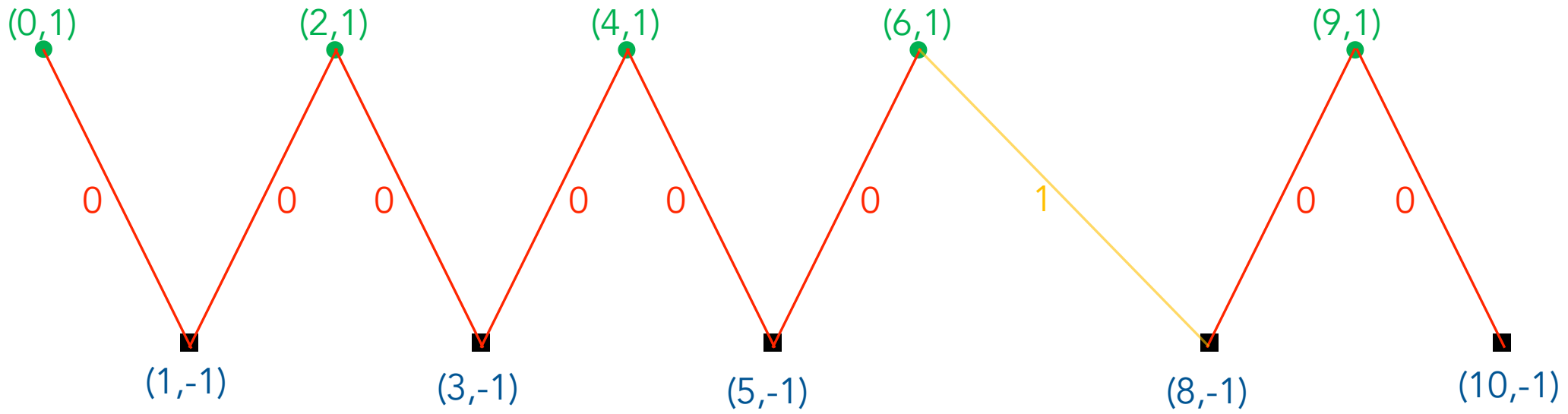
Let the adversary have a budget $\beta = 0.55$



Discrete distributions: Minimum weight matching

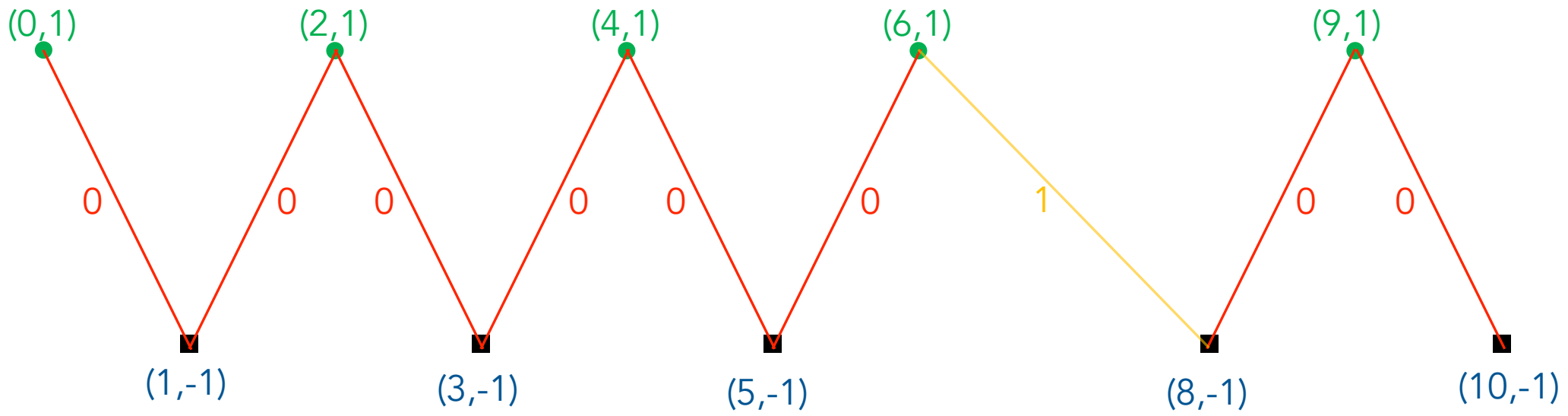


Discrete distributions: Minimum weight matching



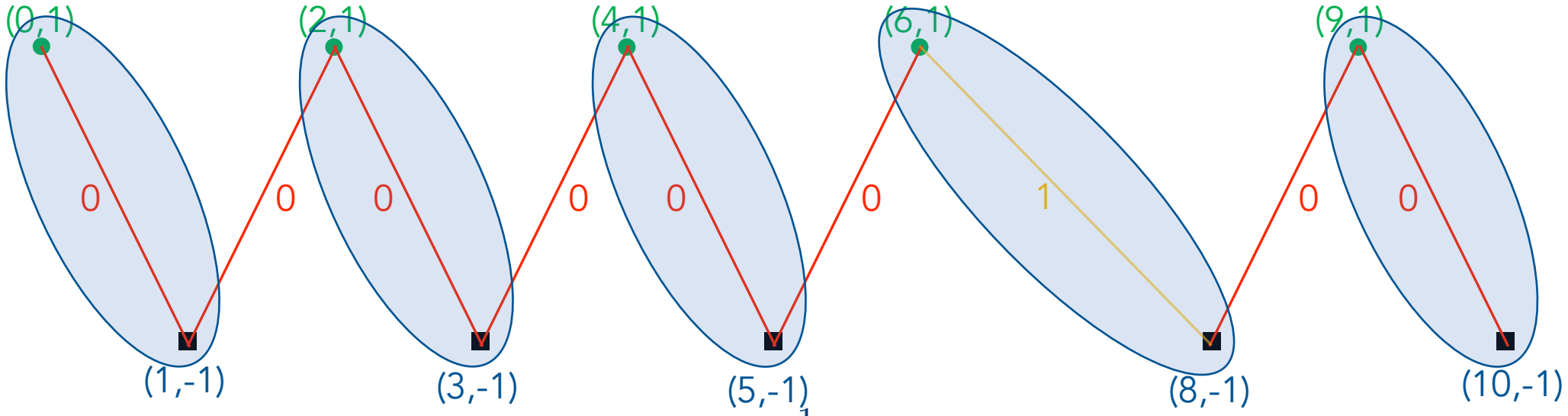
Discrete distributions: Minimum weight matching

For discrete i.i.d distributions, minimum cost is achieved at disjoint pairing of points,
i.e. **minimum weight matching**



Discrete distributions: Minimum weight matching

For discrete i.i.d distributions, minimum cost is achieved at disjoint pairing of points,
i.e. **minimum weight matching**



Cost of this matching = $\frac{1}{5}$

Loss of Optimal Classifier (with adversary) is $\frac{4}{10} = \frac{1}{2}\left(1 - \frac{1}{5}\right)$

Arbitrary distributions: Optimal Transport

Arbitrary distributions: Optimal Transport

- **Kantorovich duality** provides an alternate formulation of optimal transport in terms of **potential functions**:

$$C(P_1, P_{-1}) = \sup_{f,g} \mathbb{E}[g(X_{-1})] - \mathbb{E}[f(X_1)]$$

Arbitrary distributions: Optimal Transport

- **Kantorovich duality** provides an alternate formulation of optimal transport in terms of **potential functions**:

$$C(P_1, P_{-1}) = \sup_{f,g} \mathbb{E}[g(X_{-1})] - \mathbb{E}[f(X_1)]$$

- **Degraded classifiers** combine classification and **adversarial constraints**:

$$\tilde{h}(x) = \begin{cases} y & : N(x) \subseteq h^{-1}(y) \\ \perp & : \text{otherwise.} \end{cases}$$

Arbitrary distributions: Optimal Transport

- **Kantorovich duality** provides an alternate formulation of optimal transport in terms of **potential functions**:

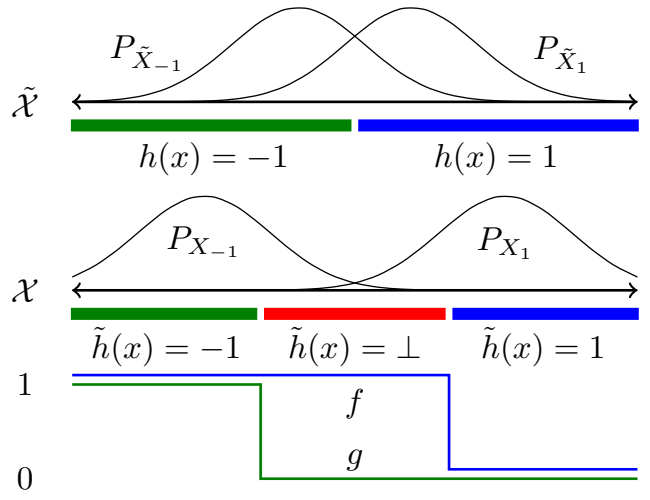
$$C(P_1, P_{-1}) = \sup_{f,g} \mathbb{E}[g(X_{-1})] - \mathbb{E}[f(X_1)]$$

- **Degraded classifiers** combine classification and **adversarial constraints**:

$$\tilde{h}(x) = \begin{cases} y & : N(x) \subseteq h^{-1}(y) \\ \perp & : \text{otherwise.} \end{cases}$$

- Following potentials are **valid for the composite cost**:

$$f(x) = 1 - \mathbf{1}[\tilde{h}(x) = 1] \qquad g(x) = \mathbf{1}[\tilde{h}(x) = -1]$$



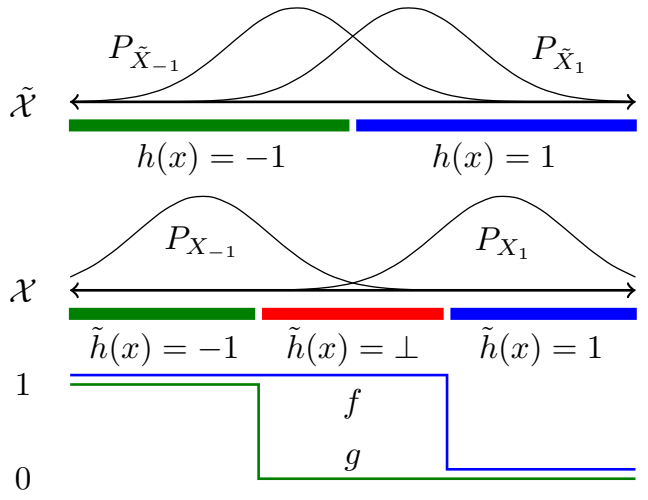
Arbitrary distributions: Optimal Transport

- **Kantorovich duality** provides an alternate formulation of optimal transport in terms of **potential functions**:

$$C(P_1, P_{-1}) = \sup_{f,g} \mathbb{E}[g(X_{-1})] - \mathbb{E}[f(X_1)]$$

- **Degraded classifiers** combine classification and **adversarial constraints**:

$$\tilde{h}(x) = \begin{cases} y & : N(x) \subseteq h^{-1}(y) \\ \perp & : \text{otherwise.} \end{cases}$$



- Following potentials are **valid for the composite cost**:

$$f(x) = 1 - \mathbf{1}[\tilde{h}(x) = 1] \qquad g(x) = \mathbf{1}[\tilde{h}(x) = -1]$$

- **Adversarial robustness** is related to **potential functions** and **degraded classifiers**:

$$\begin{aligned} 1 - L(h, P, N) &= \frac{1}{2}(\mathbb{E}[\mathbf{1}[\tilde{h}(X_1) = 1]] + \mathbb{E}[\mathbf{1}[\tilde{h}(X_{-1}) = -1]]) \\ &= \frac{1}{2}(\mathbb{E}[g(X_{-1})] + 1 - \mathbb{E}[f(X_1)]) \end{aligned}$$

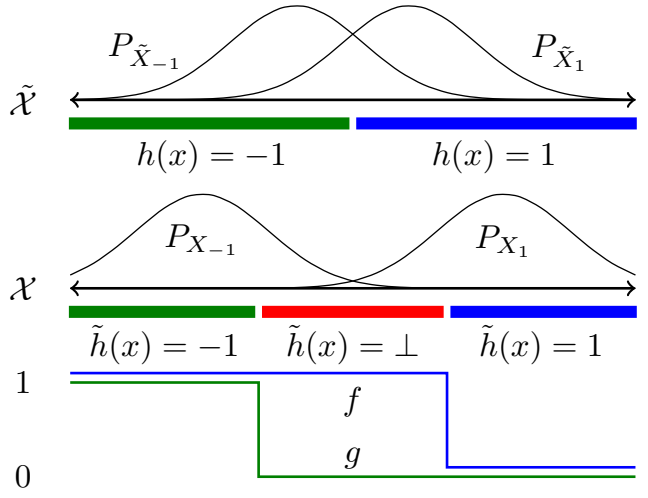
Arbitrary distributions: Optimal Transport

- **Kantorovich duality** provides an alternate formulation of optimal transport in terms of **potential functions**:

$$C(P_1, P_{-1}) = \sup_{f,g} \mathbb{E}[g(X_{-1})] - \mathbb{E}[f(X_1)]$$

- **Degraded classifiers** combine classification and **adversarial constraints**:

$$\tilde{h}(x) = \begin{cases} y & : N(x) \subseteq h^{-1}(y) \\ \perp & : \text{otherwise.} \end{cases}$$



- Following potentials are **valid for the composite cost**:

$$f(x) = 1 - \mathbf{1}[\tilde{h}(x) = 1] \qquad g(x) = \mathbf{1}[\tilde{h}(x) = -1]$$

- **Adversarial robustness** is related to **potential functions** and **degraded classifiers**:

$$\begin{aligned} 1 - L(h, P, N) &= \frac{1}{2}(\mathbb{E}[\mathbf{1}[\tilde{h}(X_1) = 1]] + \mathbb{E}[\mathbf{1}[\tilde{h}(X_{-1}) = -1]]) \\ &= \frac{1}{2}(\mathbb{E}[g(X_{-1})] + 1 - \mathbb{E}[f(X_1)]) \\ \Rightarrow 1 - L(h, P, N) &\leq \frac{1}{2}\left(1 + C(P_1, P_{-1})\right) \end{aligned}$$

Main Theorem

Main Theorem

Theorem *Adversarial constraint*

Let $\mathcal{X}$ be the space of examples and $N : \mathcal{X} \to 2^{\mathcal{X}}$ a neighborhood function on this space

Then, for any pair of distributions $P_{X_1}, P_{X_{-1}}$ on $\mathcal{X}$

$$\inf_h L(h, P, N) = \frac{1}{2}\left(1 - C(P_{X_1}, P_{X_{-1}})\right)$$

where $h : \tilde{\mathcal{X}} \to \{1, -1\}$ can be any measurable function.

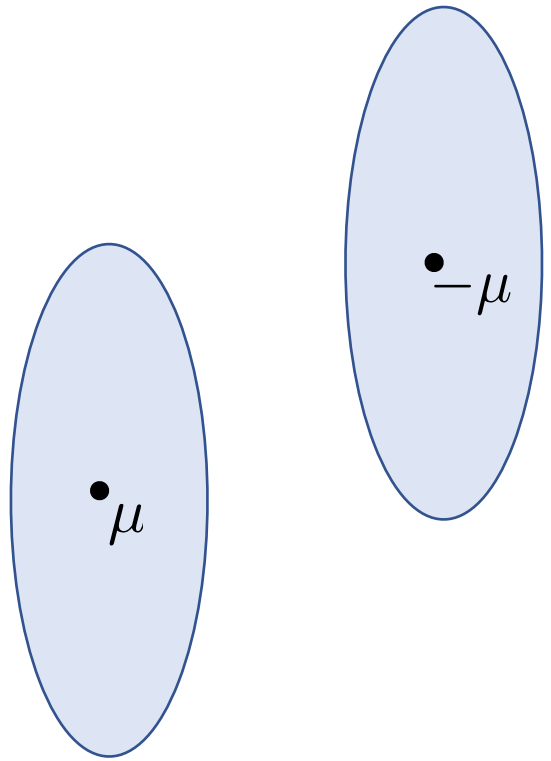
Takeaways

- Holds under very mild assumptions on space and distributions (valid for all practical cases)
- Lower bound on loss for any classifier
- Quantity on the right is easier to compute in cases of interest

Special Case: Gaussian data

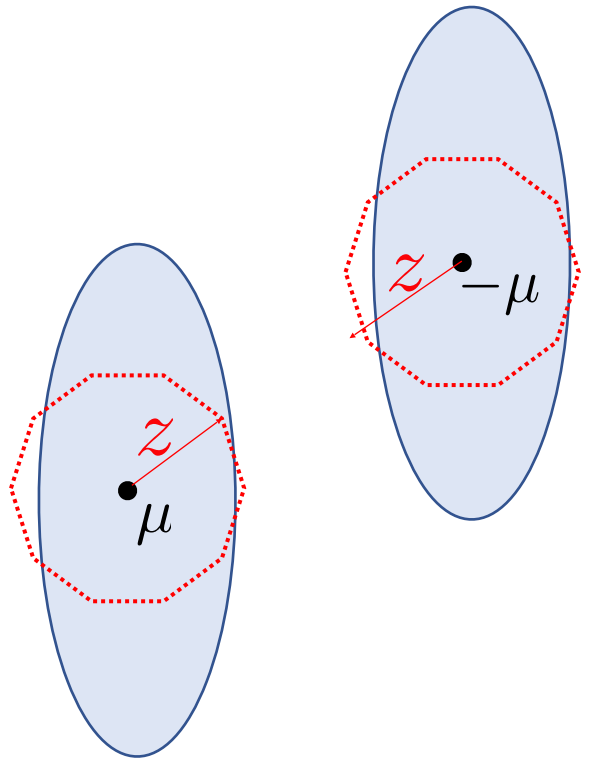
Special Case: Gaussian data

Let $P_1 = \mathcal{N}(\mu, \Sigma)$, $P_{-1} = \mathcal{N}(-\mu, \Sigma)$.



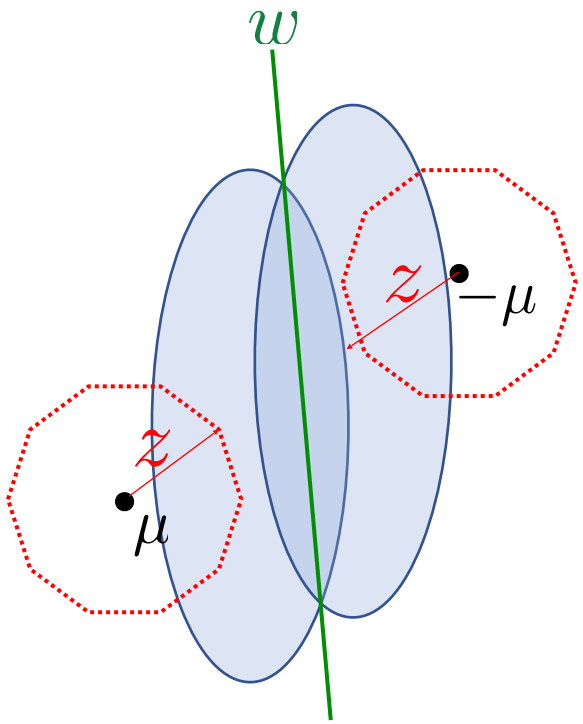
Special Case: Gaussian data

Let $P_1 = \mathcal{N}(\mu, \Sigma)$, $P_{-1} = \mathcal{N}(-\mu, \Sigma)$.
Let $N(x) = x + \beta\mathcal{B}$



Special Case: Gaussian data

Let $P_1 = \mathcal{N}(\mu, \Sigma)$, $P_{-1} = \mathcal{N}(-\mu, \Sigma)$.
Let $N(x) = x + \beta\mathcal{B}$



Special Case: Gaussian data

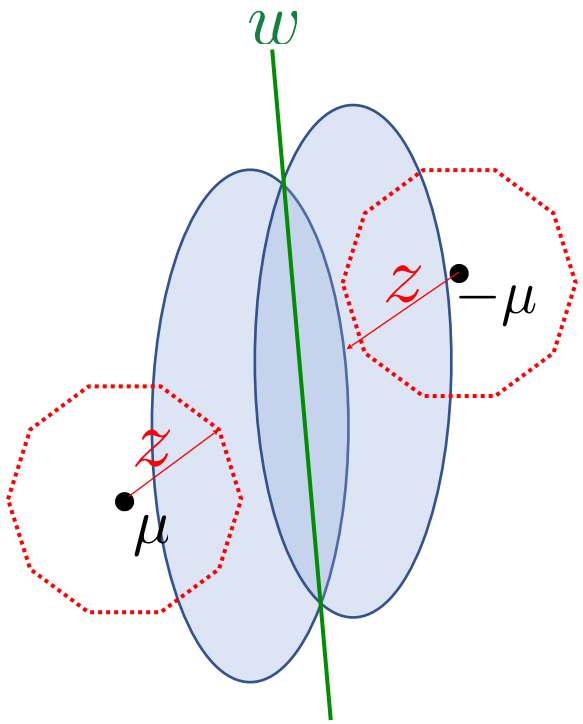
Let $P_1 = \mathcal{N}(\mu, \Sigma)$, $P_{-1} = \mathcal{N}(-\mu, \Sigma)$.
Let $N(x) = x + \beta\mathcal{B}$

Theorem

$$\inf_h L(h, P, N) = 1 - 2C(\mathcal{N}(\mu, \Sigma), \mathcal{N}(-\mu, \Sigma)) = Q(\alpha^*(\beta, \mu))$$

where $Q(\cdot)$ is the complementary cumulative distribution function of the standard normal.

$\alpha^*(\beta, \mu)$ is the solution to a convex optimization problem balancing natural and adversarial noise.



Special Case: Gaussian data

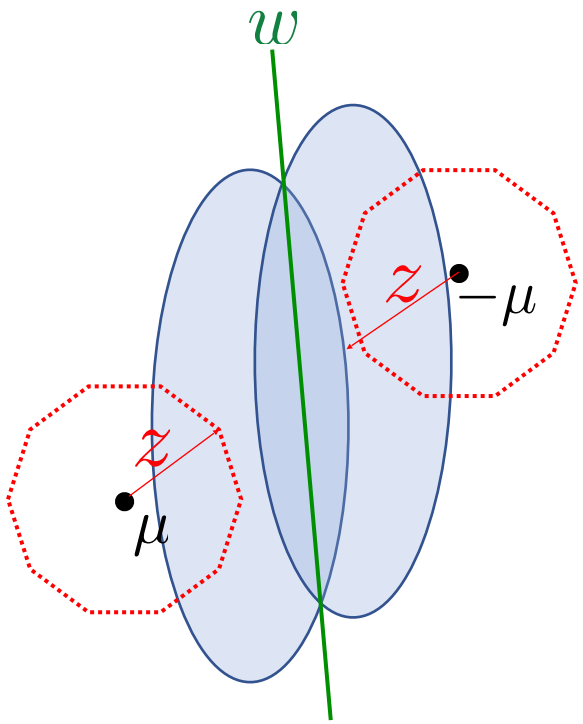
Let $P_1 = \mathcal{N}(\mu, \Sigma)$, $P_{-1} = \mathcal{N}(-\mu, \Sigma)$.
Let $N(x) = x + \beta\mathcal{B}$

Theorem

$$\inf_h L(h, P, N) = 1 - 2C(\mathcal{N}(\mu, \Sigma), \mathcal{N}(-\mu, \Sigma)) = Q(\alpha^*(\beta, \mu))$$

where $Q(\cdot)$ is the complementary cumulative distribution function of the standard normal.

$\alpha^*(\beta, \mu)$ is the solution to a convex optimization problem balancing natural and adversarial noise.



Takeaways

For convex, symmetric adversaries,

- Optimal strategy is to 'translate and pair'
- Optimal classifier is linear
- Optimal loss has a closed form

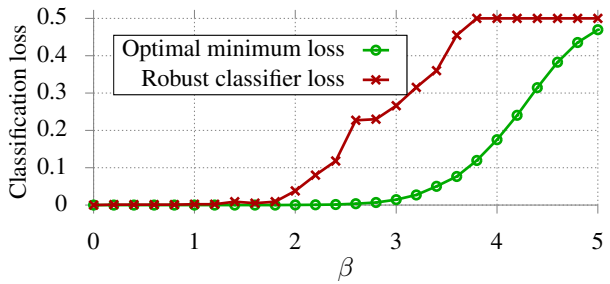
Special case: Image data

Special case: Image data

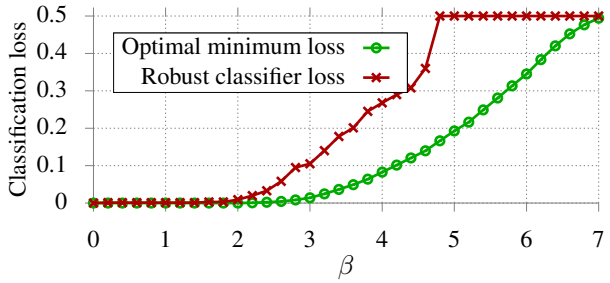
- Train robust classifier using Projected Gradient Descent (PGD)
- Compute the loss on the training data when using PGD attacks
- Optimal minimum loss computed using minimum matching on a bipartite graph

Special case: Image data

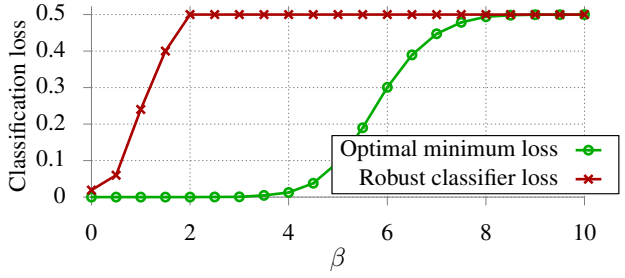
- Train robust classifier using Projected Gradient Descent (PGD)
- Compute the loss on the training data when using PGD attacks
- Optimal minimum loss computed using minimum matching on a bipartite graph



MNIST



Fashion MNIST



CIFAR-10

Pre-print at https://arxiv.org/abs/1909.12272