Training set

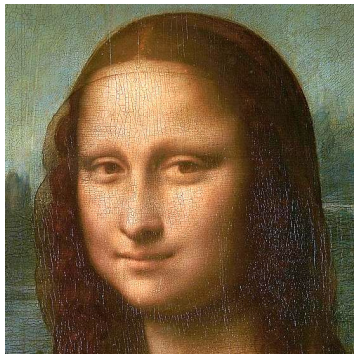
Test set

Class A:
native structure

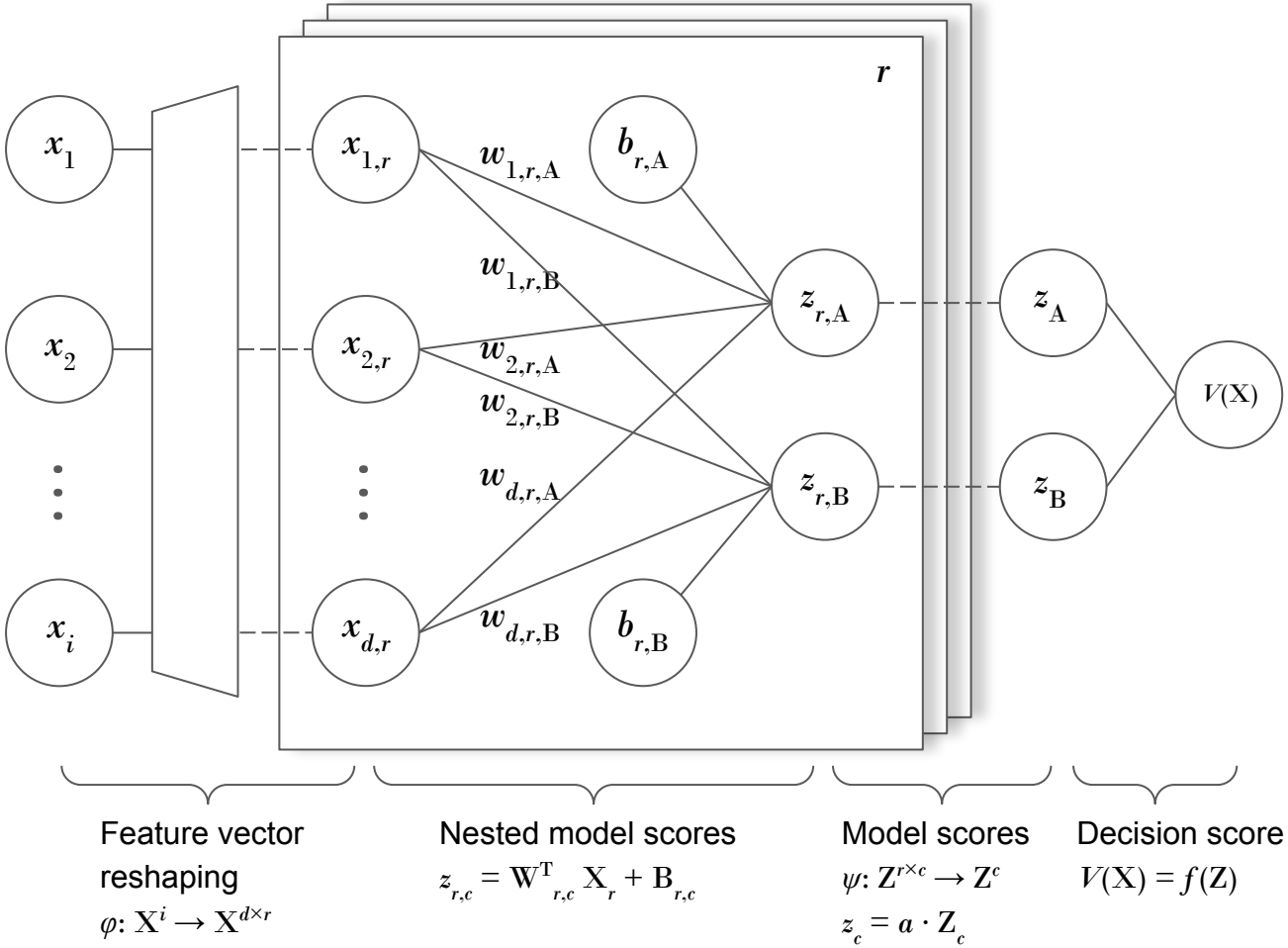
Classifier

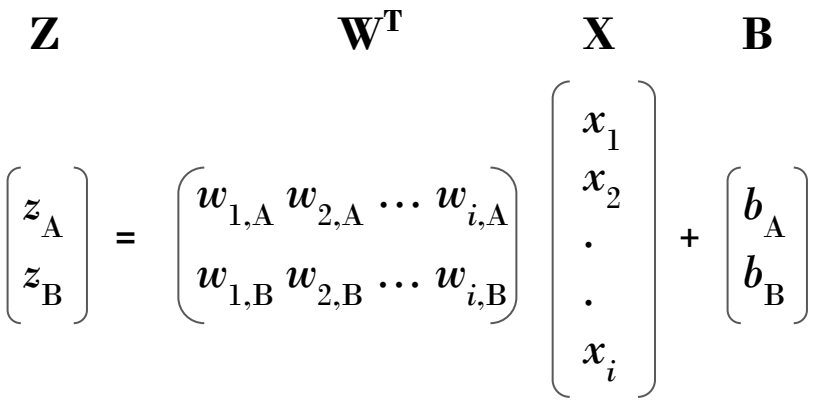
Class A:
native structure

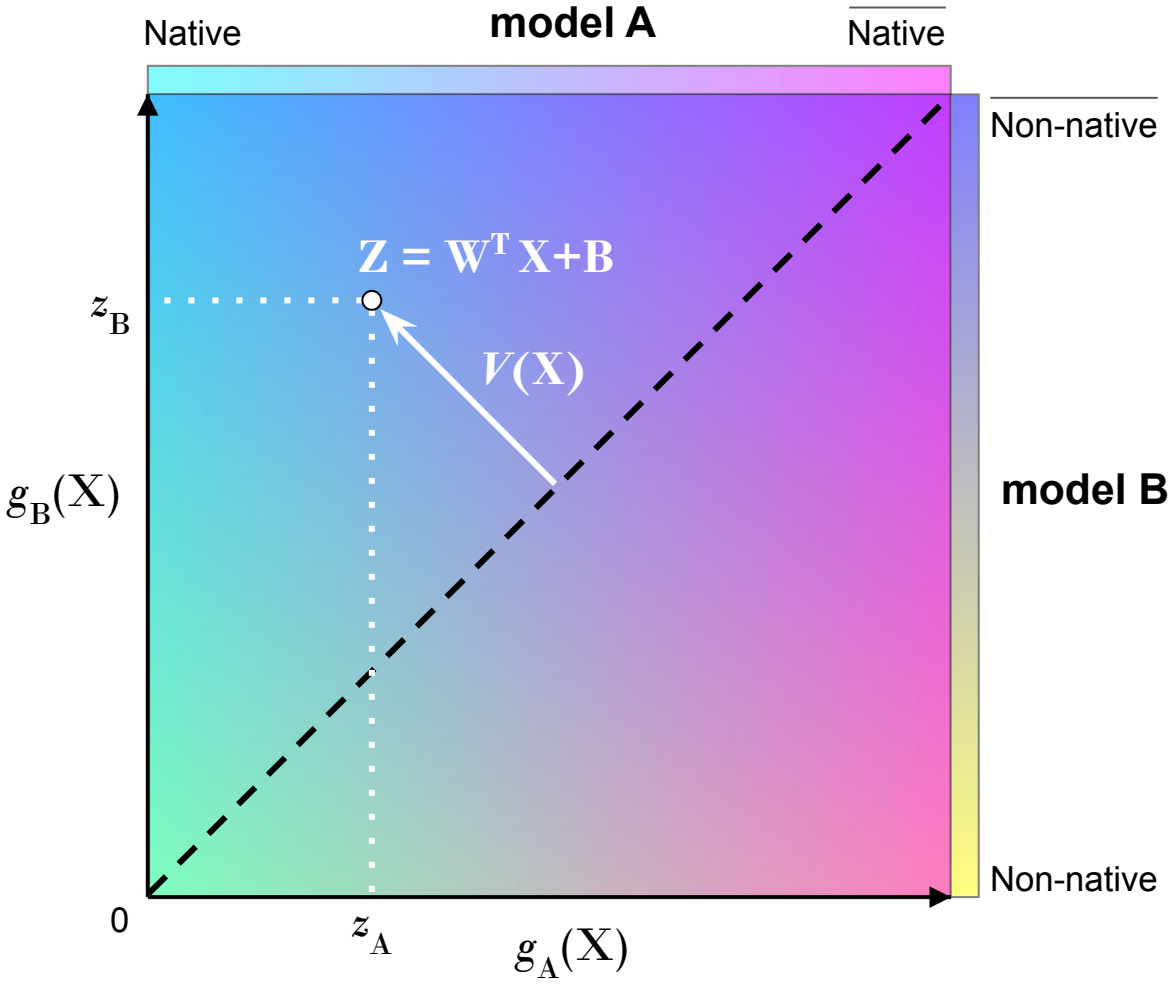
Virtual adversarial class B:
perturbed native structure



Class B:
non-native structure


r
x_1
x_2
x_i
$x_{1,r}$
$x_{2,r}$
$x_{d,r}$
$w_{1,r,\mathrm{A}}$
$w_{1,r,\mathrm{B}}$
$w_{2,r,\mathrm{A}}$
$w_{2,r,\mathrm{B}}$
$w_{d,r,\mathrm{A}}$
$w_{d,r,\mathrm{B}}$
$b_{r,\mathrm{A}}$
$b_{r,\mathrm{B}}$
$z_{r,\mathrm{A}}$
$z_{r,\mathrm{B}}$
z_{A}
z_{B}
$V(\mathrm{X})$
Feature vector reshaping
$\varphi\colon \mathbf{X}^i \to \mathbf{X}^{d\times r}$
Nested model scores
$z_{r,c} = \mathbf{W}^{\mathrm{T}}_{r,c}\,\mathbf{X}_r + \mathbf{B}_{r,c}$
Model scores
$\psi\colon \mathbf{Z}^{r\times c} \to \mathbf{Z}^c$
$z_c = a \cdot \mathbf{Z}_c$
Decision score
$V(\mathrm{X}) = f(\mathbf{Z})$


Z
W^T
X
B
z_A
z_B
=
w_1,A w_2,A ... w_i,A
w_1,B w_2,B ... w_i,B
x_1
x_2
.
.
x_i
+
b_A
b_B


Native
model A
Native
Non-native
$Z = W^T X + B$
z_B
$V(X)$
$g_B(X)$
model B
Non-native
0
z_A
$g_A(X)$