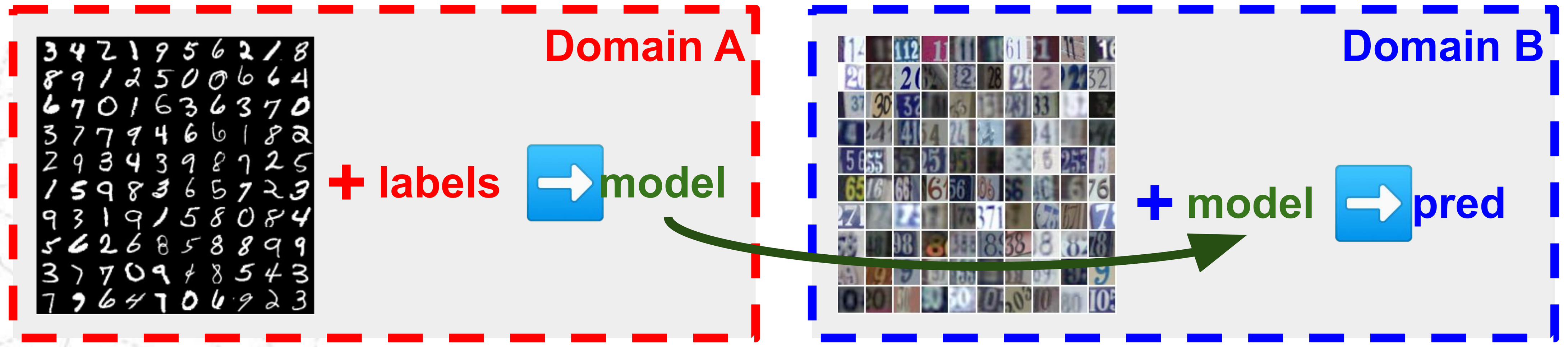
Stable Distribution Alignment Using Dual of Adversarial Distance

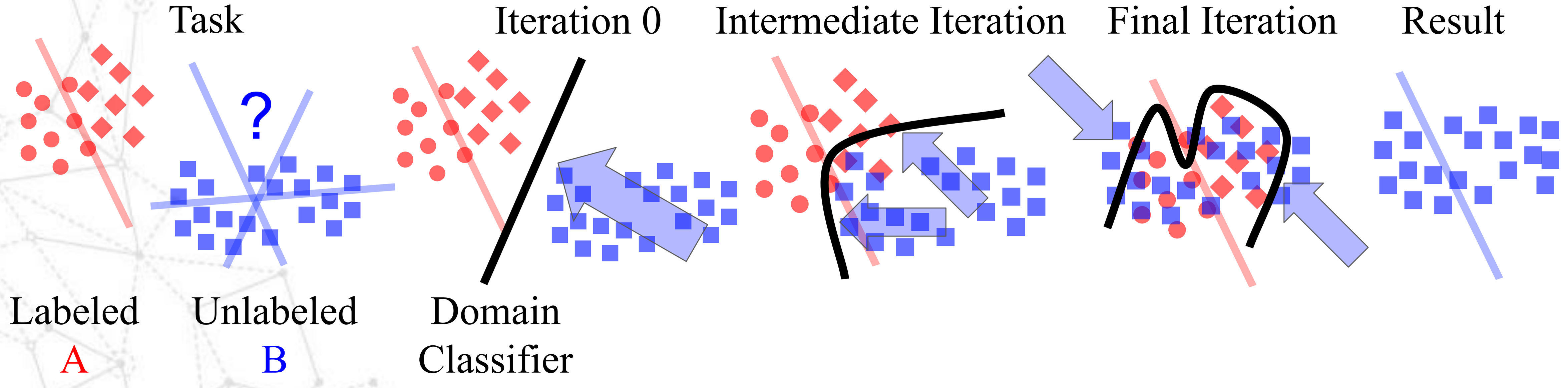
Ben Usman, Kate Saenko, Brian Kulis

BOSTON UNIVERSITY

1. Domain Adaptation



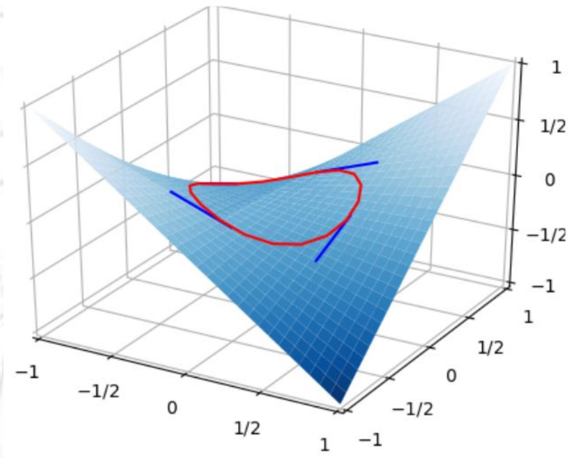
2. Adversarial Domain Adaptation produces state-of-the-art results.



Distance is measured as maximum likelihood of separating classifier:

$$\min_f d_{GAN}(A, B, f) = \min_f \max_\theta \mathcal{L}(\theta \mid A \times \{0\} \cup B_f \times \{1\})$$

3. But gradient descent is not well-suited for min-max problems, moreover we have no convergence validation procedure since optimal loss value is not necessarily smallest.



Example: Gradient descent fails to solve the saddle point problem $f(x, y) = xy$. Red line presents a trajectory of the gradient descent if vector field $(-y, x)$ is used at each iteration. Blue lines are examples of vectors from this vector field.

4. Our solution: replace min-max with min-min by dualizing logistic discriminator

$$\min_f d_{GAN}(A,B,f) = \min_f \min_{\alpha\in[0,1]^n} \mathrm{MMD}(A, B_f, \alpha) + \mathrm{H}(\alpha) = \min_f \min_{\alpha\in[0,1]^n} \langle k(A, B_f), \pi(\alpha)\rangle_F + \mathrm{H}(\alpha)$$

$$\mathrm{MMD}(A, B, [\alpha_A; \alpha_B]^T) = \frac{1}{2|A|} \sum_{0\le i,j\le |A|} K(a_i, a_j)\alpha_i^A \alpha_j^A + \frac{1}{2|B|} \sum_{0\le i,j\le |B|} K(b_i, b_j)\alpha_i^B \alpha_j^B + \frac{1}{|A||B|} \sum_{\substack{0\le i\le |A| \\ 0<i<|B|}} K(a_i, b_j)\alpha_i^A \alpha_j^B$$

Iteratively reweighted MMD and reminiscent of **Skorokhod's regularized optimal transport**

$$\pi([\alpha_A; \alpha_B]^T) = \begin{pmatrix} \alpha_A\alpha_A^T & -\alpha_A\alpha_B^T \\ -\alpha_B\alpha_A^T & \alpha_B\alpha_B^T \end{pmatrix}$$

$$k(A, B) = \begin{pmatrix} K(A, A) & K(A, B) \\ K(B, A) & K(B, B) \end{pmatrix}$$

$$H(\alpha) = \alpha^T \log(\alpha) + (1-\alpha)^T \log(1-\alpha)$$

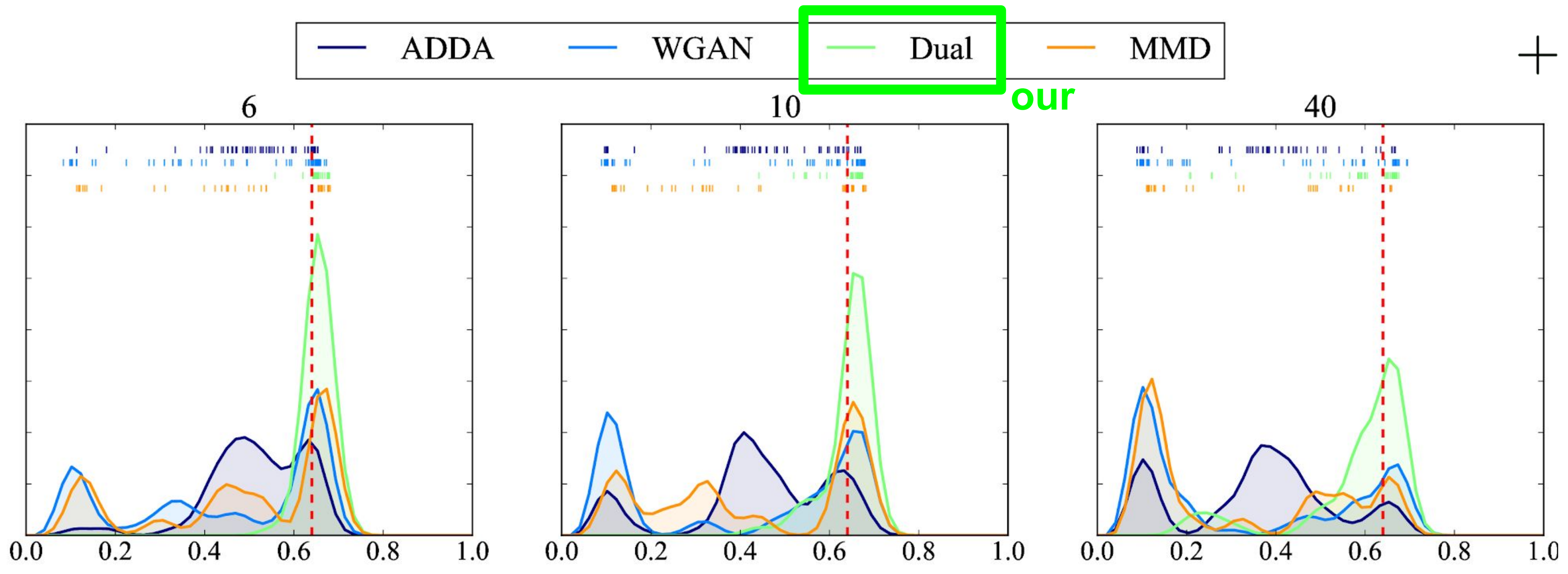


Fig 1: Kernel density estimator of **target accuracy values on last epoch.**

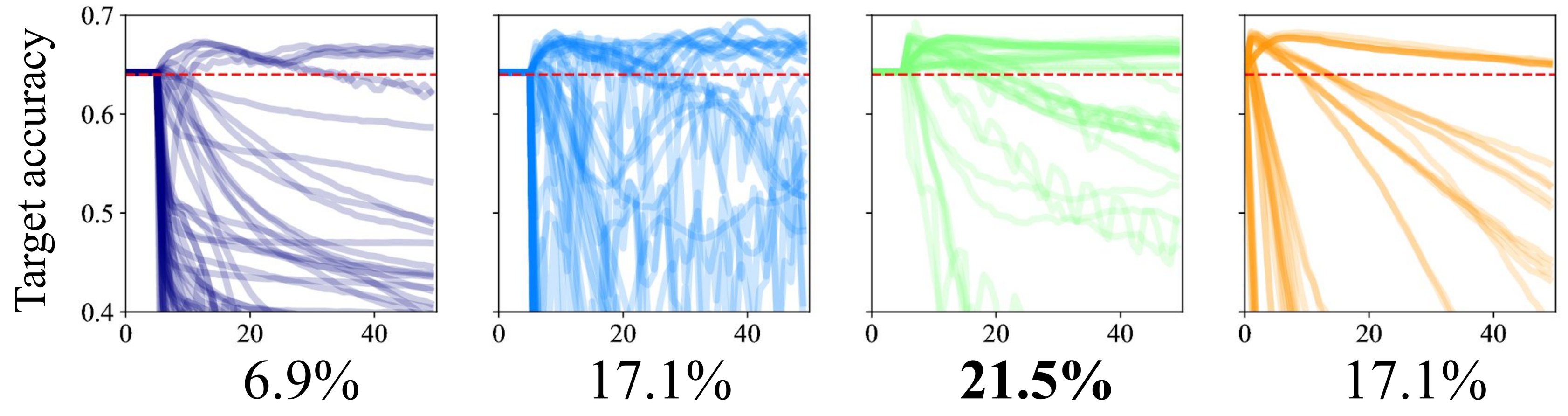


Fig 2: Evolution of target **accuracy over epochs** on MNIST-to-SVHN.

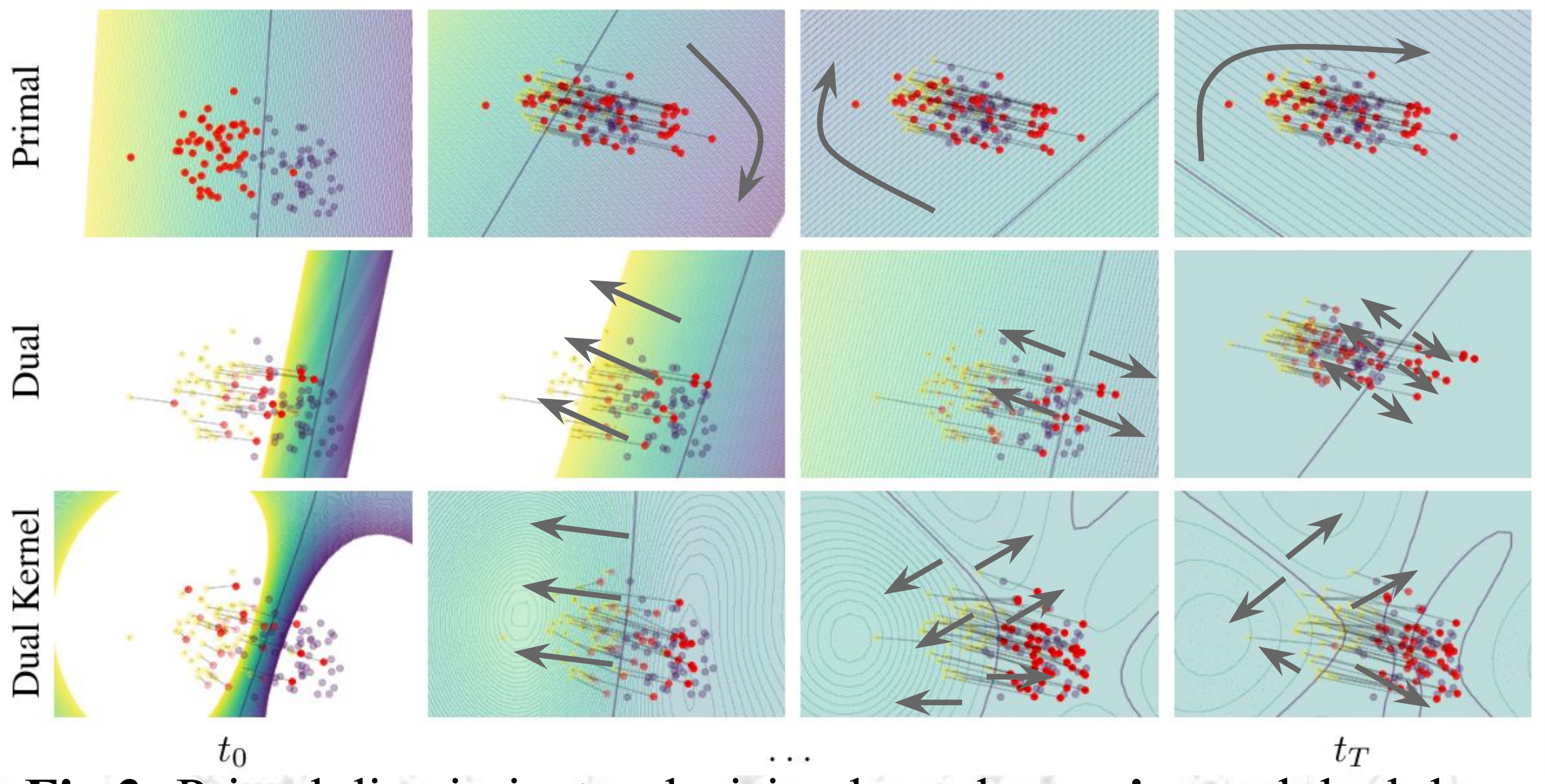


Fig 3: Primal discriminator decision boundary **spins** and dual does not.

5. Dual minimization objective makes more (21.5%) hyperparameter combinations result in models with higher then initial target accuracy.

6. On a synthetic point cloud matching task, the "primal" discriminator decision boundary *spins* around data points, whereas both linear and kernel dual approaches lead to stable solutions. We argue that this is due to discriminator being defined *implicitly* by point positions and the *alignment* matrix.

7. Future work: can we parameterize k (similarity) and π (alignment) so that they correspond to a neural discriminator?

This work was supported by NSF Award #1723379 and NSF CAREER Award #1559558

https://arxiv.org/abs/1707.04046