# KD$^2$M: A unifying framework for feature knowledge distillation

Eduardo Fernandes Montesuma

Sigma Nova
`eduardo.montesuma@sigmanova.ai`

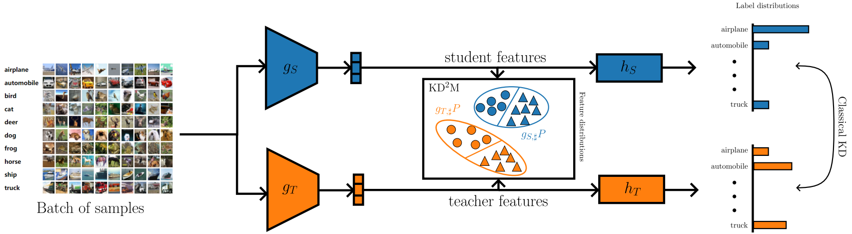
Code: https://github.com/eddardd/kddm

31 October 2025 - **GSI 2025**, Saint-Malo, Palais du Grand Large, France

# Summary

# Introduction

# Knowledge Distillation - Logits

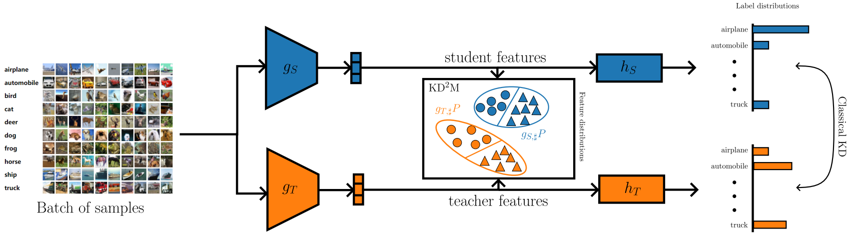



$$\theta^\star = \underset{\theta \in \Theta}{\text{argmin}} \underbrace{\underset{(\mathbf{x}^{(P)}, y^{(P)}) \sim P}{\mathbb{E}}[\mathcal{L}(y^{(P)}, h_S(g_S(\mathbf{x}^{(P)})))]}_{\text{Supervised learning objective (risk)}} + \lambda \underbrace{\mathbb{D}((h_T \circ g_T)_\sharp P, (h_S \circ g_S)_\sharp P)}_{\text{Distillation objective}},$$

- $\mathbb{D}$ is a measure of dissimilarity over $\mathcal{P}(\mathcal{Y})$.

[1] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

# Knowledge Distillation - Features (ours)




$$\theta^{\star} = \underset{\theta \in \Theta}{\text{argmin}} \underbrace{\mathbb{E}_{(\mathbf{x}^{(P)}, y^{(P)}) \sim P}[\mathcal{L}(y^{(P)}, h_S(g_S(\mathbf{x}^{(P)})))]}_{\text{Supervised learning objective (risk)}} + \lambda \underbrace{\mathbb{D}(g_{T\sharp}P, g_{S,\sharp}P)}_{\text{Distillation objective}},$$

- $\mathbb{D}$ is a measure of dissimilarity over $\mathcal{P}(\mathcal{Z} \times \mathcal{Y})$.

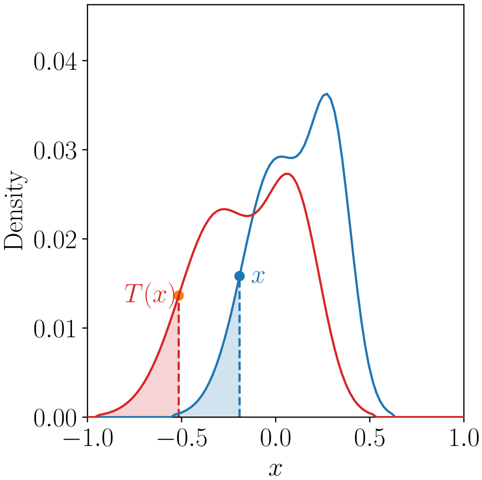
[2] Huang, Zehao, and Naiyan Wang. "Like what you like: Knowledge distill via neuron selectivity transfer." arXiv preprint arXiv:1707.01219 (2017).

[3] Lohit, Suhas, and Michael Jones. "Model compression using optimal transport." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

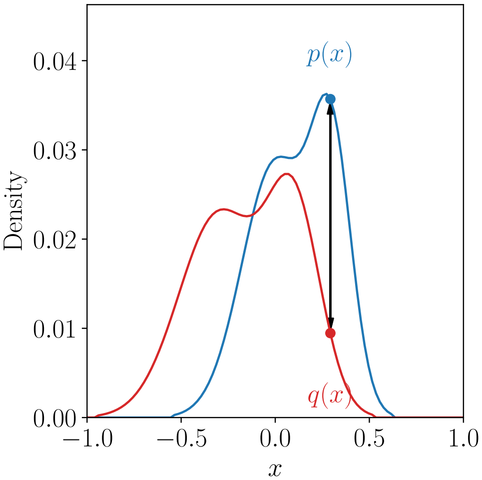
[4] Lv, Jiaming, Haoyuan Yang, and Peihua Li. "Wasserstein distance rivals kullback-leibler divergence for knowledge distillation." Advances in Neural Information Processing Systems 37 (2024): 65445-65475.

# Probability Metrics

# Probability Metrics

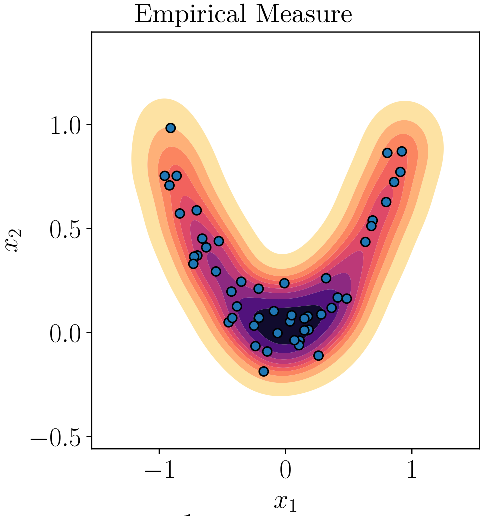



$$\mathbb{W}_2(P,Q)^2 = \inf_{\gamma \in \Gamma(P,Q)} \int_{\mathcal{Z}} \int_{\mathcal{Z}} \|z - z'\|_2^2 \, d\gamma(z,z')$$

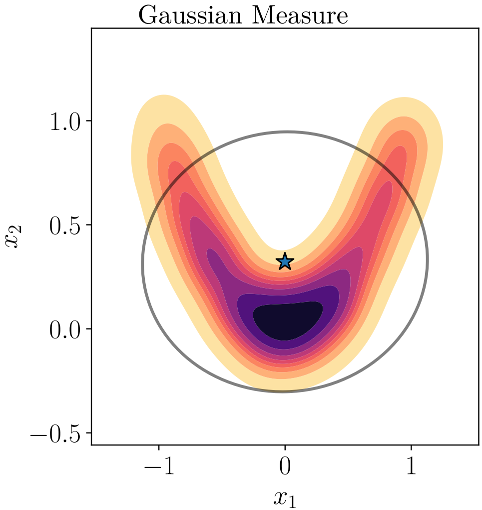



$$\mathbb{KL}(P,Q) = \int_{\mathcal{Z}} \log \frac{P(z)}{Q(z)} \, dP(z)$$

# Probability Metrics - Finite Approximations

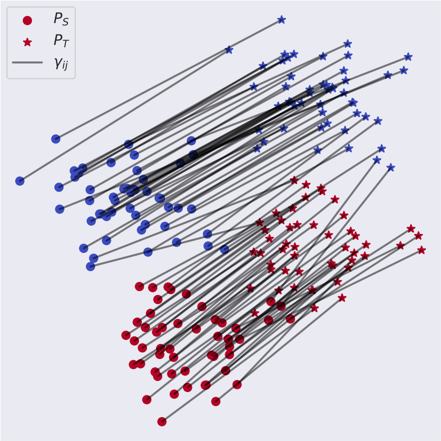


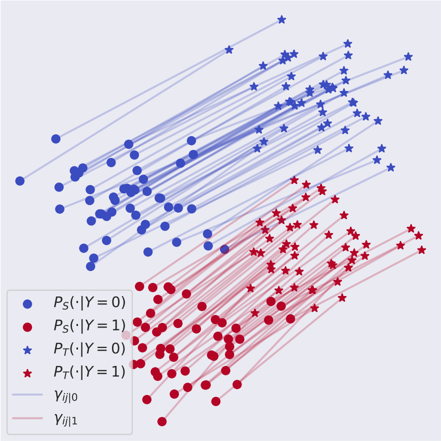

$$P(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\mathbf{z} - \mathbf{z}_i^{(P)})$$




$$P(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mu^{(P)}, \Sigma^{(P)})$$

# The Empirical Case



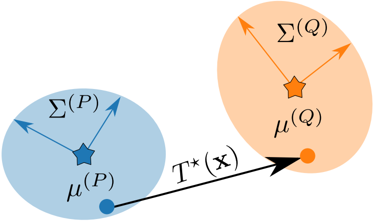




$$\mathbb{W}_2(\hat{P},\hat{Q})^2 = \textstyle\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\|\mathbf{z}_i^{(P)} - z_j^{(Q)}\|_2^2 \qquad \mathbb{CW}_2(\hat{P},\hat{Q})^2 = \frac{1}{n_c}\textstyle\sum_{y=1}^{n_c}\mathbb{W}_2(\hat{P}(Z|Y=y),\hat{Q}(Z|Y=y))^2$$

$$\mathbb{JW}_2(\hat{P}_S,\hat{P}_T)^2 = \min_{\gamma\in\Gamma(\hat{P},\hat{Q})}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}(\|\mathbf{z}_i^{(P_S)} - \mathbf{z}_j^{(P_T)}\|^2 + \beta\mathcal{L}(h(\mathbf{z}_i^{(P_S)}), h(\mathbf{z}_j^{(P_T)})),$$

# The Gaussian Case

Gaussian Measure




Wasserstein:

$$\mathbb{W}_2(P,Q)^2 = \|\mu^{(P)} - \mu^{(Q)}\|_2^2 - \mathbb{B}(\Sigma^{(P)}, \Sigma^{(Q)}),$$

$$\mathbb{B}(\Sigma^{(P)}, \Sigma^{(Q)}) = \mathrm{Tr}\left(\Sigma_P + \Sigma_Q - 2(S_P \Sigma_Q S_P)^{1/2}\right),$$

where $S_P = \mathrm{sqrtm}(\Sigma_P)$ (resp. $Q$).

Kullback-Leibler:

$$\mathbb{KL}(P|Q) = \frac{1}{2}\left(\mathrm{Tr}((\Sigma^{(Q)})^{-1}\Sigma^{(P)}) + (\mu^{(Q)} - \mu^{(P)})^T (\Sigma^{(Q)})^{-1}(\mu^{(Q)} - \mu^{(P)}) - d + \log\left(\frac{\det(\Sigma^{(Q)})}{\det(\Sigma^{(P)})}\right)\right),$$

# KDDM: Knowledge Distillation through Distribution Matching

# KDDM - Algorithm

**Algorithm 0:** Training step of KD$^2$M

**Function** `training_on_minibatch`($\{\mathbf{x}_i^{(P)}, y_i^{(P)}\}_{i=1}^{n}, \lambda$)

// Forward pass - Student

$\mathbf{Z}^{(P_S)} \leftarrow \{g_S(\mathbf{x}_i^{(P)})\}_{i=1}^{n}$, $\hat{\mathbf{Y}}^{(P_S)} \leftarrow \{h_S(\mathbf{z}_i^{(P_S)})\}$

// Classification loss - Student

$\mathcal{L}_c \leftarrow -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{n_c} y_{ic}^{(P)} \log \hat{y}_{ic}^{(P_S)}$

// Forward pass - Teacher

$\mathbf{Z}^{(P_T)} \leftarrow \{g_T(\mathbf{x}_i^{(P)})\}_{i=1}^{n}$, $\hat{\mathbf{Y}}^{(P_T)} \leftarrow \{h_T(\mathbf{z}_i^{(P_T)})\}$

// Feature distillation loss

$\mathcal{L}_d \leftarrow \text{compute\_distribution\_distance}(\mathbf{Z}^{(P_S)}, \mathbf{Z}^{(P_T)}, \mathbf{Y}^{(P)}, \hat{\mathbf{Y}}^{(P_S)}, \hat{\mathbf{Y}}^{(P_T)})$

**return** $\mathcal{L}_c + \lambda\mathcal{L}_d$

# Theoretical Results

**Definition.** (Error) Given a $P \in \mathcal{P}(\mathcal{X})$, a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and a ground truth $f_0 : \mathcal{X} \to \mathcal{Y}$, the generalization error of $f$ is,

$$\mathcal{R}_P(f) = \mathbb{E}_{x \sim P}[\mathcal{L}(f(x), f_0(x))]$$

**Lemma** (Wasserstein bound) Let $\mathcal{Z} \subset \mathbb{R}^d$ be separable. Let $P_S, P_T \in \mathcal{P}(\mathcal{Z})$. Assume $c(\mathbf{z}, \mathbf{z}') = \|\mathbf{z} - \mathbf{z}'\|_{\mathcal{H}_k}$, where $\mathcal{H}_k$ is a reproducing kernel Hilbert space with kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ induced by $\phi : \mathcal{Z} \to \mathcal{H}_k$. Assume that $\mathcal{L}_{h,h'}(\mathbf{z}) = |h(\mathbf{z}) - h'(\mathbf{z})|$, and that $k$ is squared root integrable with respect $P_S$ and $P_T$, and $0 \leq k(\mathbf{z}, \mathbf{z}') \leq K, \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}$. Assuming $\|\mathcal{L}\|_{\mathcal{H}_k} \leq 1$,

$$|\mathcal{R}_{P_S}(h) - \mathcal{R}_{P_T}(h)| \leq \mathbb{W}_2(P_S, P_T).$$

**Theorem** (ours) Under the same conditions of Lemma 1, let $P \in \mathcal{P}(\mathcal{X})$ be a fixed distribution. Let $g_S$ and $g_T$ be two measurable mappings from $\mathcal{X}$ to a latent space $\mathcal{Z} \subset \mathbb{R}^d$, such that $\|g_S\|_{L_2(P)} < \infty$ and $\|g_T\|_{L_2(P)} < \infty$. Define $P_S = g_{S,\sharp}P$ and $P_T = g_{T,\sharp}P$, then,

$$|\mathcal{R}_{P_S}(h) - \mathcal{R}_{P_T}(h)| \leq \|g_S - g_T\|_{L_2(P)}.$$

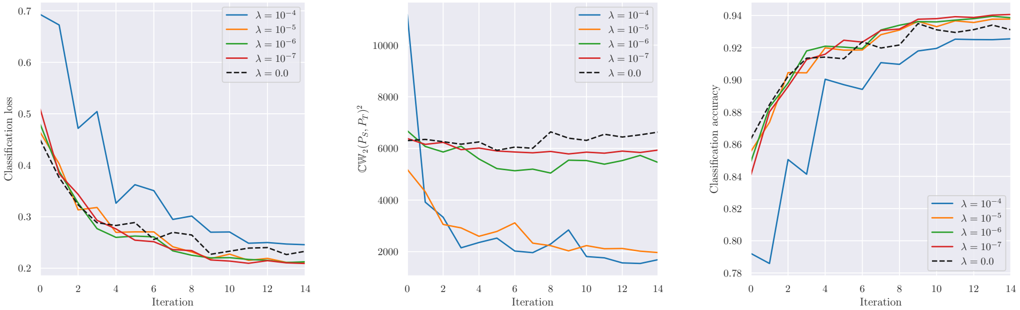
[5] Redko, Ievgen, Amaury Habrard, and Marc Sebban. "Theoretical analysis of domain adaptation with optimal transport." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 2017.

# Empirical Results

Table: Classification accuracy (in %) of KDDM for different distribution metrics on computer vision benchmarks. Distances are either over empirical (E) or Gaussian (G) approximations.

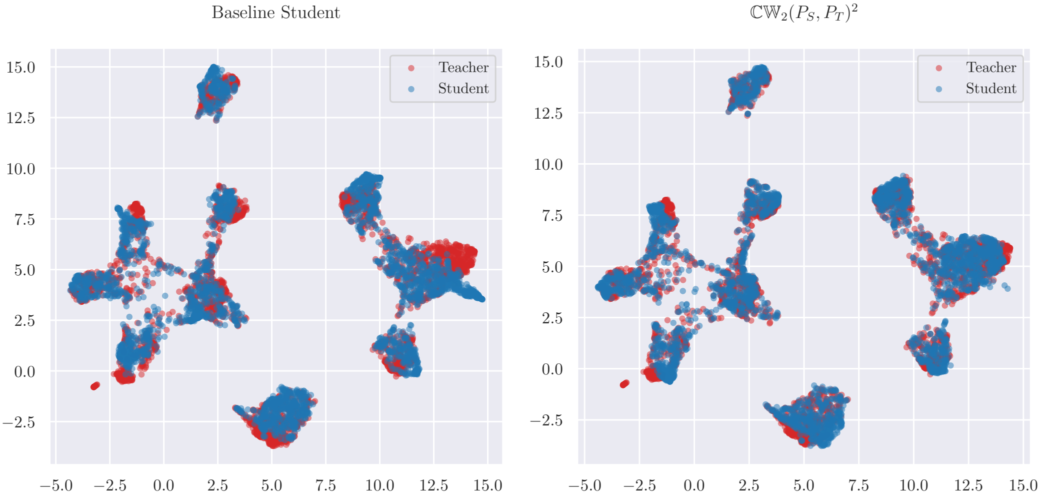
| Method | SVHN | CIFAR-10 | CIFAR-100 | Avg. |
|---|---|---|---|---|
| Student | ResNet18 | ResNet18 | ResNet18 | – |
| Teacher | ResNet34 | ResNet34 | ResNet34 | – |
| Student | 93.10 | 85.11 | 56.66 | 78.29 |
| Teacher | 94.41 | 86.98 | 62.21 | 81.20 |
| $\mathbb{W}_2$ (E) | 94.00 | 86.45 | 61.07 | 80.51 |
| $\mathbb{CW}_2$ (E) | **94.06** | 86.54 | **61.47** | **80.69** |
| $\mathbb{JW}_2$ (E) | 94.00 | **86.60** | 61.07 | 80.55 |
| $\mathbb{W}_2$ (G) | 93.94 | 86.63 | 60.68 | 80.41 |
| $\mathbb{CW}_2$ (G) | 93.95 | 86.25 | 61.43 | 80.54 |
| $\mathbb{KL}$ (G) | 94.05 | 86.44 | 60.66 | 80.38 |

# Empirical Results




**Tradeoff:** Distillation loss generally helps. Too strong $\lambda$ hurts task (classification) performance

# Empirical Results



# Conclusion

# Empirical Results

Our work,

- Aggregates previous works under a common framework,
- Derives a new theoretical understanding for Knowledge Distillation
- Using distances over $\mathcal{P}(\mathcal{Z} \times \mathcal{Y})$ *generally works better.*

Future works,

- Distillation on larger scale settings,
- Refine theoretical results
- Design new probability metircs/divergences for distillation
- Can we go beyond a distance between $g_S$ and $g_T$?

Our code is available at: `https://github.com/eddardd/kddm`