Being Properly Improper

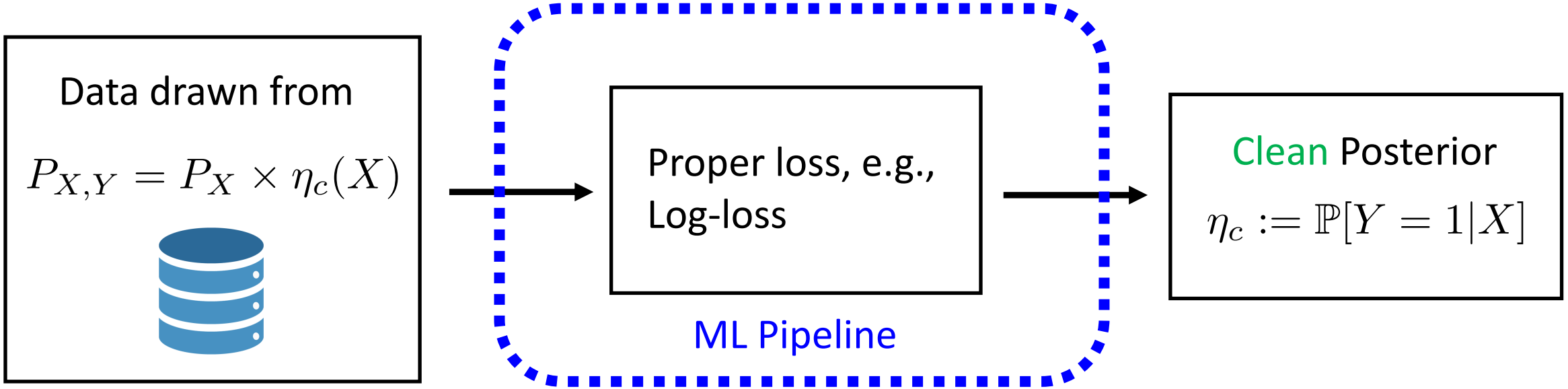
Tyler Sypherd

Richard Nock

Lalitha Sankar

Properness

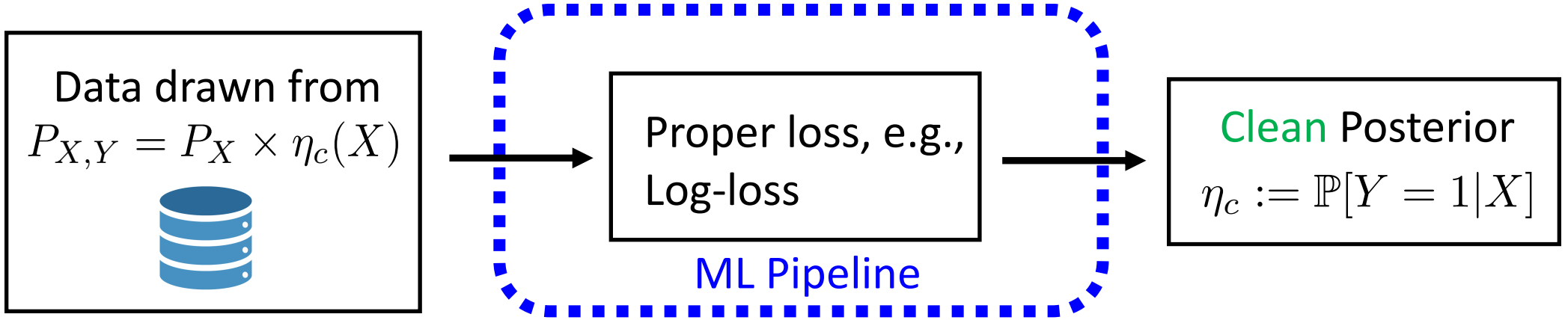
- Proper loss estimates the true posterior



Reid and Williamson. "Composite binary losses." (2010).
Reid and Williamson. "Information, divergence and risk for binary experiments." (2011).

Properness

- Proper loss estimates the true posterior



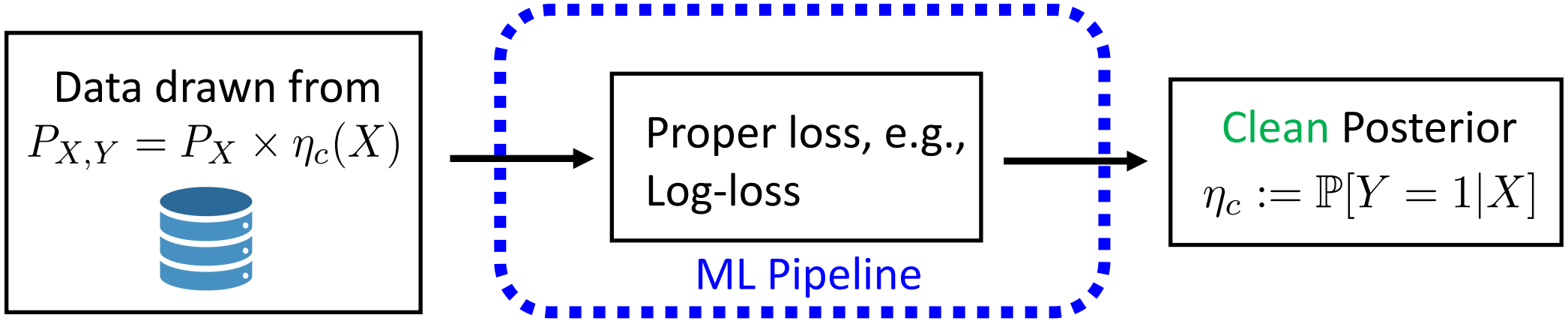
- Conditional risk of ℓ of estimated posterior $\eta \in [0, 1]$ wrt ground truth $\eta_c \in [0, 1]$

$$L(\eta, \eta_c) := \mathbb{E}_{Y \sim \mathrm{Ber}(\eta_c)}[\ell(Y, \eta)] = \eta_c \ell_1(\eta) + (1 - \eta_c)\ell_{-1}(\eta)$$

Reid and Williamson. "Composite binary losses." (2010).
Reid and Williamson. "Information, divergence and risk for binary experiments." (2011).

Properness

- Proper loss estimates the true posterior



- Conditional risk of ℓ of estimated posterior $\eta \in [0, 1]$ wrt ground truth $\eta_c \in [0, 1]$

$$L(\eta, \eta_c) := \mathbb{E}_{Y \sim \mathrm{Ber}(\eta_c)}[\ell(Y, \eta)] = \eta_c \ell_1(\eta) + (1 - \eta_c)\ell_{-1}(\eta)$$

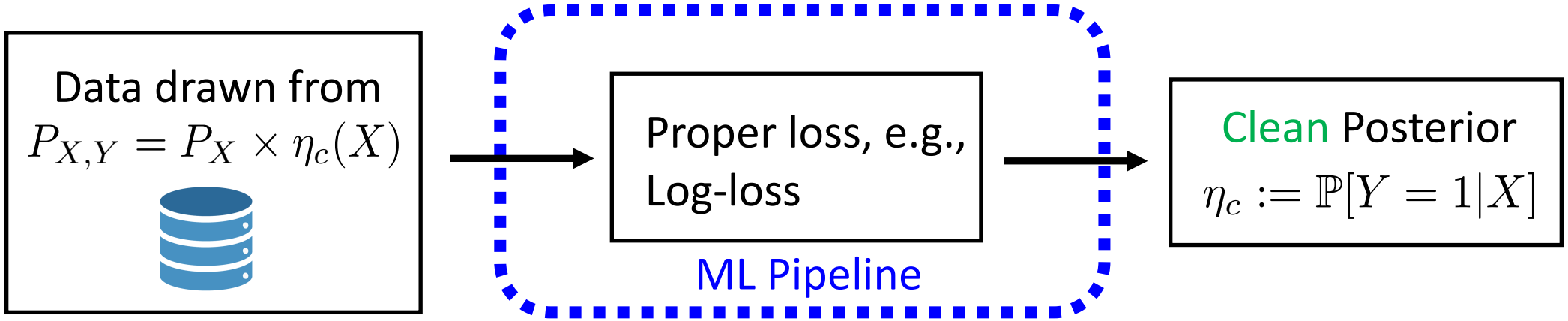
- Proper loss:

$$\arg\min_{\eta} L(\eta, \eta_c) = \eta_c$$

Reid and Williamson. "Composite binary losses." (2010).
Reid and Williamson. "Information, divergence and risk for binary experiments." (2011).

Properness

- Proper loss estimates the true posterior



- Conditional risk of ℓ of estimated posterior $\eta \in [0, 1]$ wrt ground truth $\eta_c \in [0, 1]$

$$L(\eta, \eta_c) := \mathbb{E}_{Y \sim \mathrm{Ber}(\eta_c)}[\ell(Y, \eta)] = \eta_c \ell_1(\eta) + (1 - \eta_c)\ell_{-1}(\eta)$$

- Example: log-loss

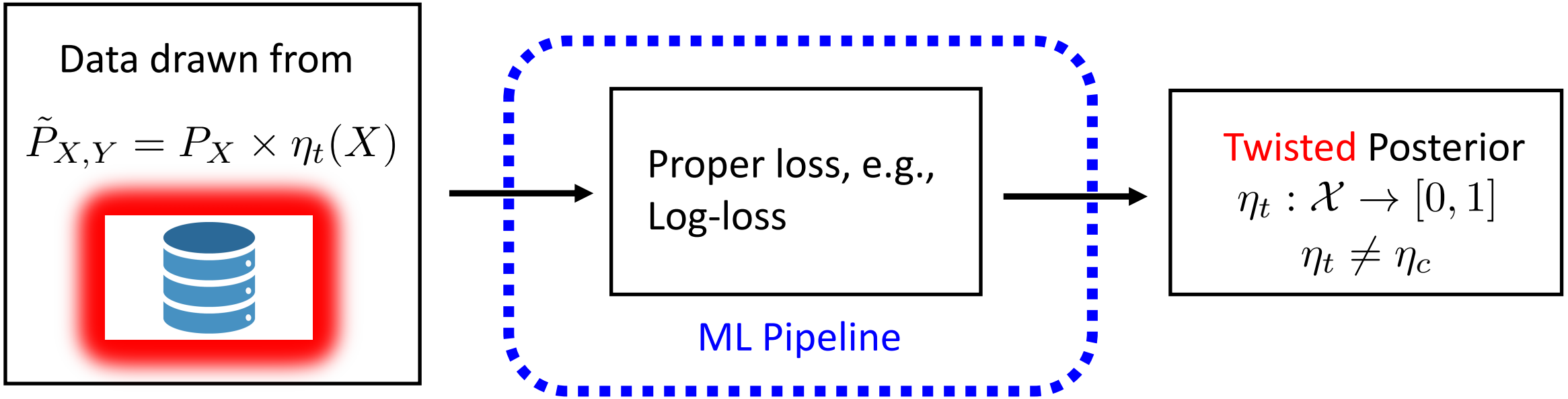
$$\ell_1(\eta) := -\log \eta$$
$$\ell_{-1}(\eta) := -\log (1 - \eta)$$

Reid and Williamson. "Composite binary losses." (2010).
Reid and Williamson. "Information, divergence and risk for binary experiments." (2011).

Cross-entropy loss is built with log-loss

The Problem with Properness

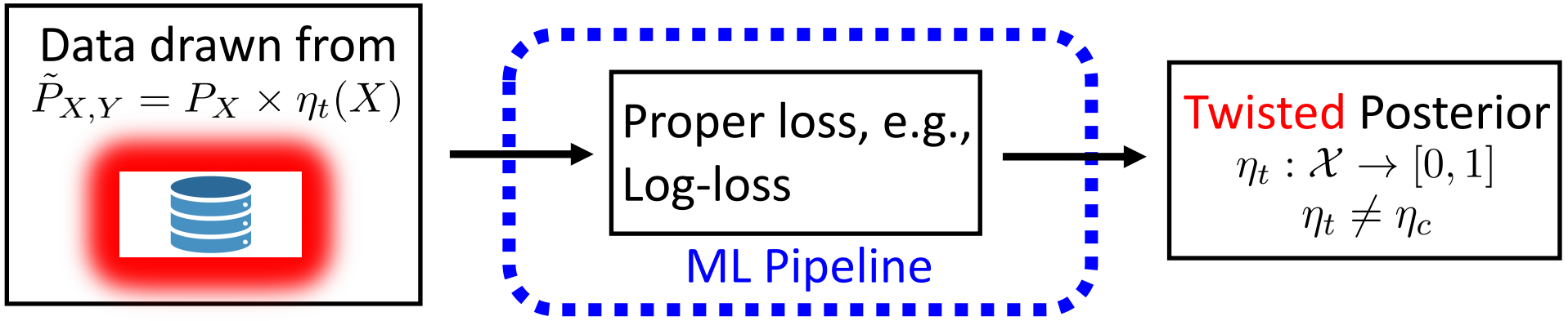
- What happens if the data is twisted? (e.g., label/feature/adversarial noise)



Sypherd, Nock, Sankar, ”Being Properly Improper”. https://arxiv.org/abs/2106.09920.

The Problem with Properness

- What happens if the data is twisted? (e.g., label/feature/adversarial noise)



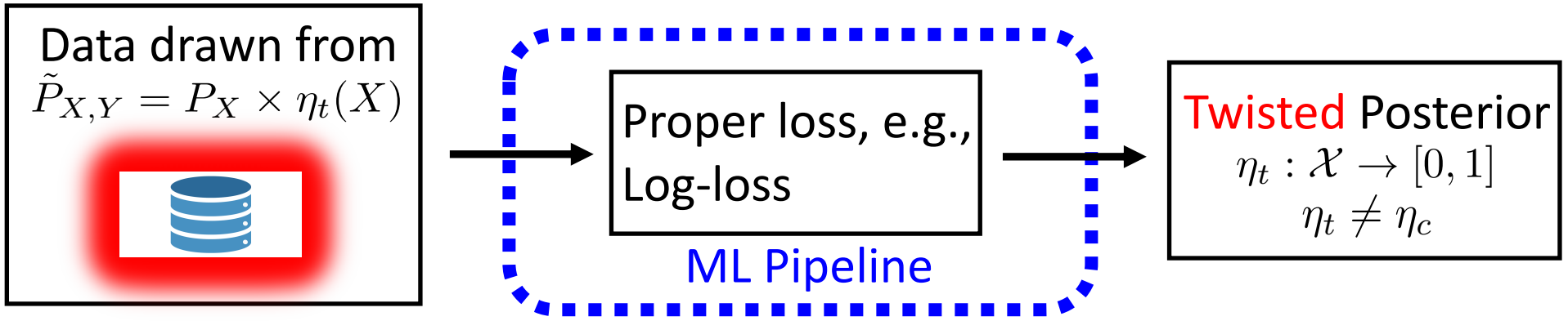
- Proper loss minimizes and returns:

$$\arg\min_{\eta} L(\eta, \eta_t) = \arg\min_{\eta} \eta_t \ell_1(\eta) + (1 - \eta_t)\ell_{-1}(\eta) = \eta_t$$

Sypherd, Nock, Sankar, ”Being Properly Improper”. https://arxiv.org/abs/2106.09920.

The Problem with Properness

- What happens if the data is twisted? (e.g., label/feature/adversarial noise)



- Proper loss minimizes and returns:

$$\arg\min_{\eta} L(\eta, \eta_t) = \arg\min_{\eta} \eta_t \ell_1(\eta) + (1-\eta_t)\ell_{-1}(\eta) = \eta_t$$

- A twist $\eta_c \to \eta_t$ refers to a general mapping (label/feature/adversarial noise)

Sypherd, Nock, Sankar, "Being Properly Improper". https://arxiv.org/abs/2106.09920.

Twist-Proper Losses: Untwisting Twisted Posterior

- A loss is twist-proper iff for any twist $\eta_c \to \eta_t$, there exist hyperparameter(s) β^*, s.t.,

$$\arg\min_{\eta} L^{\beta^*}(\eta, \eta_t) = \arg\min_{\eta} \eta_t \ell_1^{\beta^*}(\eta) + (1-\eta_t)\ell_{-1}^{\beta^*}(\eta) = \eta_c$$

- This means for every $x \in \mathcal{X}$, a twist-proper loss has hyperparameter(s) β^* that untwist the twisted posterior into the clean posterior

1. Lin, Goyal, Girshick, He, and Dollár (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).
2. Castells, Weinzaepfel, and Revaud (2020). Superloss: A generic loss for robust curriculum learning. *Advances in Neural Information Processing Systems*, *33*, 4308-4319.
3. Sypherd, Diaz, Cava, Dasarathy, Kairouz and Sankar (2022). "A Tunable Loss Function for Robust Classification: Calibration, Landscape, and Generalization," in *IEEE Transactions on Information Theory.*

Twist-Proper Losses: Untwisting Twisted Posterior

- A loss is twist-proper iff for any twist $\eta_c \to \eta_t$, there exist hyperparameter(s) β^*, s.t.,

$$\arg\min_{\eta} L^{\beta^*}(\eta, \eta_t) = \arg\min_{\eta} \eta_t \ell_1^{\beta^*}(\eta) + (1-\eta_t)\ell_{-1}^{\beta^*}(\eta) = \eta_c$$

- This means for every $x \in \mathcal{X}$, a twist-proper loss has hyperparameter(s) β^* that untwist the twisted posterior into the clean posterior

- Not vacuous: Focal loss and Super Loss are not twist-proper

- Alpha-loss (generalizes log-loss) studied by Sypherd et al. (2022) is twist-proper

1. Lin, Goyal, Girshick, He, and Dollár (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).
2. Castells, Weinzaepfel, and Revaud (2020). Superloss: A generic loss for robust curriculum learning. *Advances in Neural Information Processing Systems*, *33*, 4308-4319.
3. Sypherd, Diaz, Cava, Dasarathy, Kairouz and Sankar (2022). "A Tunable Loss Function for Robust Classification: Calibration, Landscape, and Generalization," in *IEEE Transactions on Information Theory.*

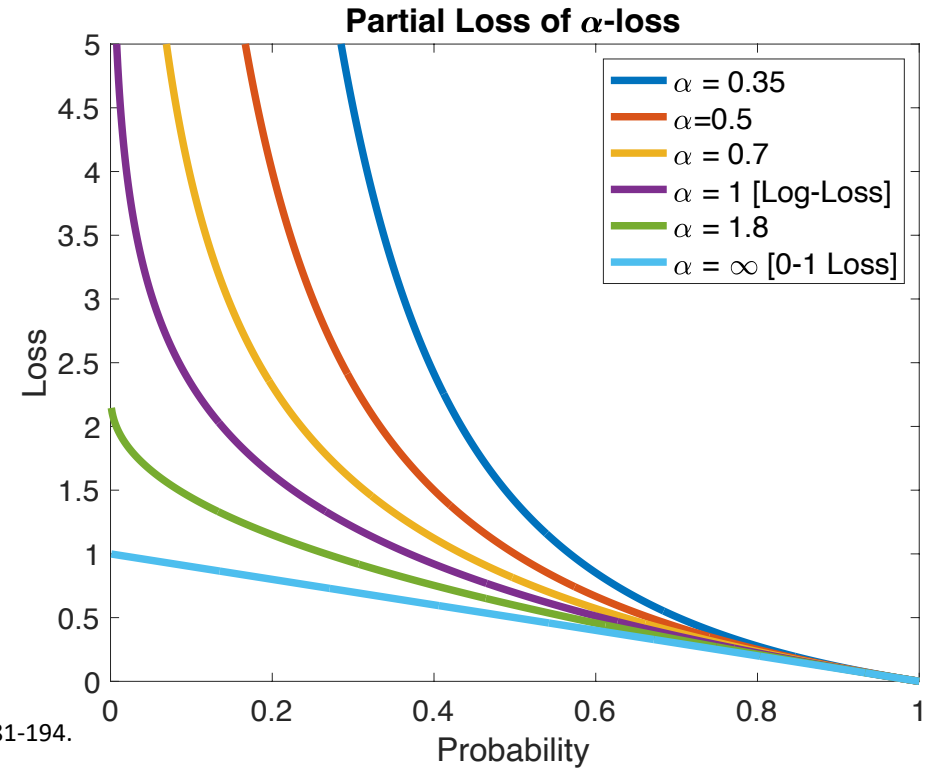
α-loss is Twist-Proper

- For $\alpha \geq 0$, partial losses $\ell_1^\alpha(\eta) := \ell_{-1}^\alpha(1-\eta), \forall \eta \in [0,1]$,

$$\ell_1^\alpha(\eta) := \frac{\alpha}{\alpha - 1}\left(1 - \eta^{\frac{\alpha-1}{\alpha}}\right)$$

- And, by continuity:

$$\ell_1^0(\eta) := \infty \quad \ell_1^1(\eta) := -\log \eta \quad \ell_1^\infty(\eta) := 1 - \eta$$



Liao, Kosut, Sankar, Calmon. "A tunable measure for information leakage." *2018 IEEE ISIT*.
S. Arimoto. "Information-theoretical considerations on estimation problems." Information and control 19.3 (1971): 181-194.

α-loss is Twist-Proper

- For $\alpha \geq 0$, partial losses $\ell_1^\alpha(\eta) := \ell_{-1}^\alpha(1-\eta), \forall \eta \in [0,1]$,

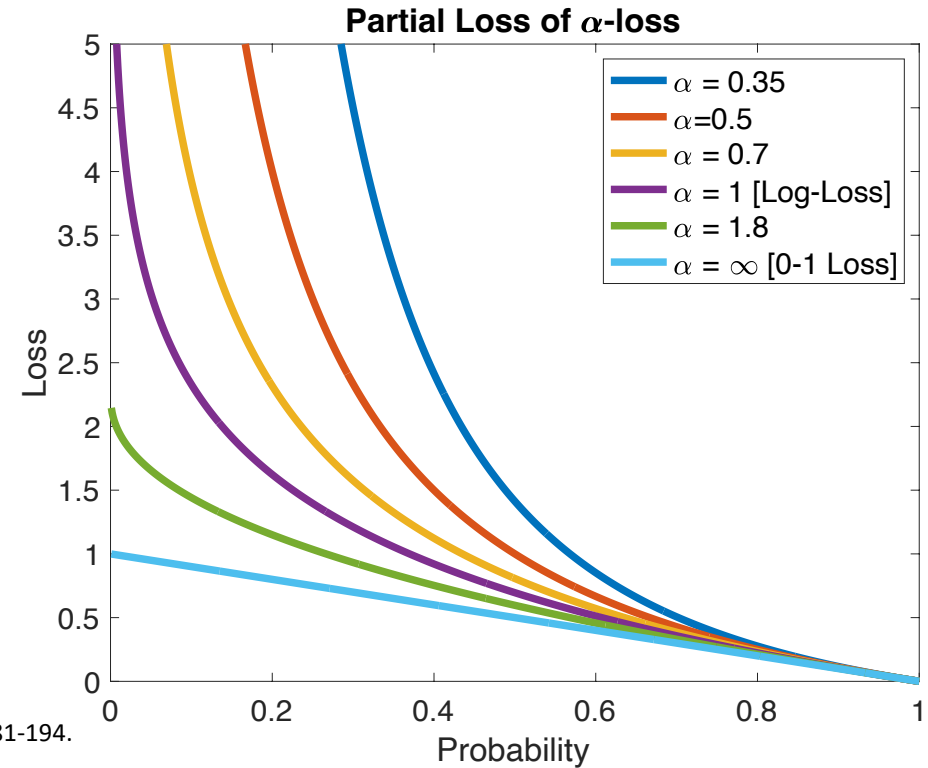
$$\ell_1^\alpha(\eta) := \frac{\alpha}{\alpha - 1}\left(1 - \eta^{\frac{\alpha-1}{\alpha}}\right)$$

- And, by continuity:

$$\ell_1^0(\eta) := \infty \quad \ell_1^1(\eta) := -\log \eta \quad \ell_1^\infty(\eta) := 1 - \eta$$

- Twist-proper (via alpha-tilted distribution):

$$\arg\min_{\eta} L^\alpha(\eta, \eta_t) = \frac{\eta_t^\alpha}{\eta_t^\alpha + (1-\eta_t)^\alpha}$$



Liao, Kosut, Sankar, Calmon. "A tunable measure for information leakage." *2018 IEEE ISIT*.
S. Arimoto. "Information-theoretical considerations on estimation problems." Information and control 19.3 (1971): 181-194.

Untwisting $\mathcal{X}$ with a fixed $\alpha_0 > 1$

Theorem: For any strictly Bayes blunting twist $\eta_c \to \eta_t$, there exists a fixed $\alpha_0 > 1$ which induces the following ordering:

$$D_{\mathrm{KL}}(\eta_c, \eta_t; 1) > D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha_0)$$

- Strictly Bayes blunting twist $(\eta_c < \eta_t < 1/2) \vee (\eta_c > \eta_t > 1/2)$

Sypherd T., Nock R., Sankar L., ”Being Properly Improper”. https://arxiv.org/abs/2106.09920.

Untwisting $\mathcal{X}$ with a fixed $\alpha_0 > 1$

Theorem: For any strictly Bayes blunting twist $\eta_c \to \eta_t$, there exists a fixed $\alpha_0 > 1$ which induces the following ordering:

$$D_{\mathrm{KL}}(\eta_c, \eta_t; 1) > D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha_0)$$

- Strictly Bayes blunting twist $(\eta_c < \eta_t < 1/2) \vee (\eta_c > \eta_t > 1/2)$
- This answers yes to untwisting the feature space with a fixed $\alpha_0 > 1$
- SLN $\eta_t = p(1 - \eta_c) + (1 - p)\eta_c$ is a strictly Bayes blunting twist for $0 < p < 1/2$

Sypherd T., Nock R., Sankar L., ”Being Properly Improper”. https://arxiv.org/abs/2106.09920.

Pseudo-Inverse-Link Boost

- α-loss has been investigated in logistic regression and neural networks
- Novel boosting algorithm in order to boost hyperparameterized loss

Algorithm 1 PILBOOST

Input sample $\mathcal{S} = \{(\boldsymbol{x}_i, y_i), i = 1, 2, ..., m\}$, number of iterations T, $a_f > 0$, PIL $\widetilde{f}$;
Step 1 : let $\boldsymbol{\beta} \leftarrow \mathbf{0}$; // first classifier, $H_{\mathbf{0}} = 0$
Step 2 : **for** $t = 1, 2, ..., T$
  Step 2.1 : **for** $i = 1, 2, ..., m$, let $w_i \leftarrow \widetilde{f}(-y_i H_{\boldsymbol{\beta}}(x_i))$ // PIL weights
  Step 2.2 : let $j \leftarrow \text{WL}(\mathcal{S}, \boldsymbol{w})$
  Step 2.3 : let $\eta_j \leftarrow (1/m) \cdot \sum_i w_i y_i h_j(\boldsymbol{x}_i)$
  Step 2.4 : let $\beta_j \leftarrow \beta_j + a_f \eta_j$
Return $H_{\boldsymbol{\beta}}$.

Sypherd T., Nock R., Sankar L., "Being Properly Improper". https://arxiv.org/abs/2106.09920.

Pseudo-Inverse-Link Boost

- α-loss has been investigated in logistic regression and neural networks
- Novel boosting algorithm in order to boost hyperparameterized loss

Algorithm 1 PILBOOST

Input sample $\mathcal{S} = \{(\boldsymbol{x}_i, y_i), i = 1, 2, ..., m\}$, number of iterations T, $a_f > 0$, PIL $\widetilde{f}$;
Step 1 : let $\boldsymbol{\beta} \leftarrow \mathbf{0}$; // first classifier, $H_{\mathbf{0}} = 0$
Step 2 : **for** $t = 1, 2, ..., T$
  Step 2.1 : **for** $i = 1, 2, ..., m$, let $w_i \leftarrow \widetilde{f}(-y_i H_{\boldsymbol{\beta}}(x_i))$ // PIL weights
  Step 2.2 : let $j \leftarrow \text{WL}(\mathcal{S}, \boldsymbol{w})$
  Step 2.3 : let $\eta_j \leftarrow (1/m) \cdot \sum_i w_i y_i h_j(\boldsymbol{x}_i)$
  Step 2.4 : let $\beta_j \leftarrow \beta_j + a_f \eta_j$
Return $H_{\boldsymbol{\beta}}$.

- PILBoost yields boosting compliant rates
- Gains over AdaBoost and XGBoost on twisted data (label/feature/adversarial noise)

Sypherd T., Nock R., Sankar L., "Being Properly Improper". https://arxiv.org/abs/2106.09920.

Pseudo-Inverse-Link Boost

- α-loss has been investigated in logistic regression and neural networks
- Novel boosting algorithm in order to boost hyperparameterized loss

Algorithm 1 PILBOOST

Input sample $\mathcal{S} = \{(\boldsymbol{x}_i, y_i), i = 1, 2, ..., m\}$, number of iterations T, $a_f > 0$, PIL $\widetilde{f}$;
Step 1 : let $\boldsymbol{\beta} \leftarrow \mathbf{0}$; // first classifier, $H_{\mathbf{0}} = 0$
Step 2 : **for** $t = 1, 2, ..., T$
  Step 2.1 : **for** $i = 1, 2, ..., m$, let $w_i \leftarrow \widetilde{f}(-y_i H_{\boldsymbol{\beta}}(x_i))$ // PIL weights
  Step 2.2 : let $j \leftarrow \text{WL}(\mathcal{S}, \boldsymbol{w})$
  Step 2.3 : let $\eta_j \leftarrow (1/m) \cdot \sum_i w_i y_i h_j(\boldsymbol{x}_i)$
  Step 2.4 : let $\beta_j \leftarrow \beta_j + a_f \eta_j$
Return $H_{\boldsymbol{\beta}}$.

- PILBoost yields boosting compliant rates
- Gains over AdaBoost and XGBoost on twisted data (label/feature/adversarial noise)

Thank you!