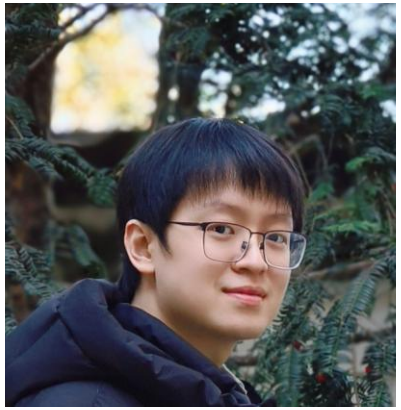
Convolutional Fenchel–Young Loss

Convex Smooth Losses with Linear Surrogate Regret




Yuzhou Cao [1] Han Bao [2] Lei Feng [3] Bo An [1,4]

[1]Nanyang Technological University [2]the Institute of Statistical Mathematics [3]Southeast University [4]Skywork AI

Learning vs. Evaluation

Evaluation stage: Performance measuring with **discrete** loss function ℓ *w.r.t.* **discrete (finite)** prediction space $\widehat{\mathcal{Y}}$ $\left(|\widehat{\mathcal{Y}}| < +\infty\right)$

$$\min_{h:\mathcal{X}\to\widehat{\mathcal{Y}}} R_\ell(h) = \mathbb{E}_{X,Y}[\ell(h(X),Y)]$$

Learning vs. Evaluation

Evaluation stage: Performance measuring with **discrete** loss function ℓ *w.r.t.* **discrete (finite)** prediction space $\widehat{\mathcal{Y}}$ $\left(|\widehat{\mathcal{Y}}| < +\infty\right)$

$$\min_{h:\mathcal{X}\to\widehat{\mathcal{Y}}} R_\ell(h) = \mathbb{E}_{X,Y}[\ell(h(X),Y)]$$

Hard to optimize

Learning vs. Evaluation

Evaluation stage: Performance measuring with **discrete** loss function ℓ *w.r.t.* **discrete (finite)** prediction space $\widehat{\mathcal{Y}}$ $(|\widehat{\mathcal{Y}}| < +\infty)$.

$$\min_{h:\mathcal{X}\to\widehat{\mathcal{Y}}} R_\ell(h) = \mathbb{E}_{X,Y}[\ell(h(X),Y)]$$

Hard to optimize

Learning stage: Minimization of (expected) **continuous** loss function ϕ *w.r.t.* **continuous** prediction space $\mathbb{R}^d$.

$$\min_{f:\mathcal{X}\to\mathbb{R}^d} R_\phi(f) = \mathbb{E}_{X,Y}[\phi(f(X),Y)]$$

Continuous **Surrogate**

Learning vs. Evaluation

Evaluation stage: Performance measuring with **discrete** loss function ℓ *w.r.t.* **discrete (finite)** prediction space $\widehat{\mathcal{Y}}$ $(|\widehat{\mathcal{Y}}| < +\infty)$.

$$\min_{h:\mathcal{X}\to\widehat{\mathcal{Y}}} R_\ell(h) = \mathbb{E}_{X,Y}[\ell(h(X),Y)]$$

Hard to optimize

Learning stage: Minimization of (expected) **continuous** loss function ϕ *w.r.t.* **continuous** prediction space $\mathbb{R}^d$.

$$\min_{f:\mathcal{X}\to\mathbb{R}^d} R_\phi(f) = \mathbb{E}_{X,Y}[\phi(f(X),Y)]$$

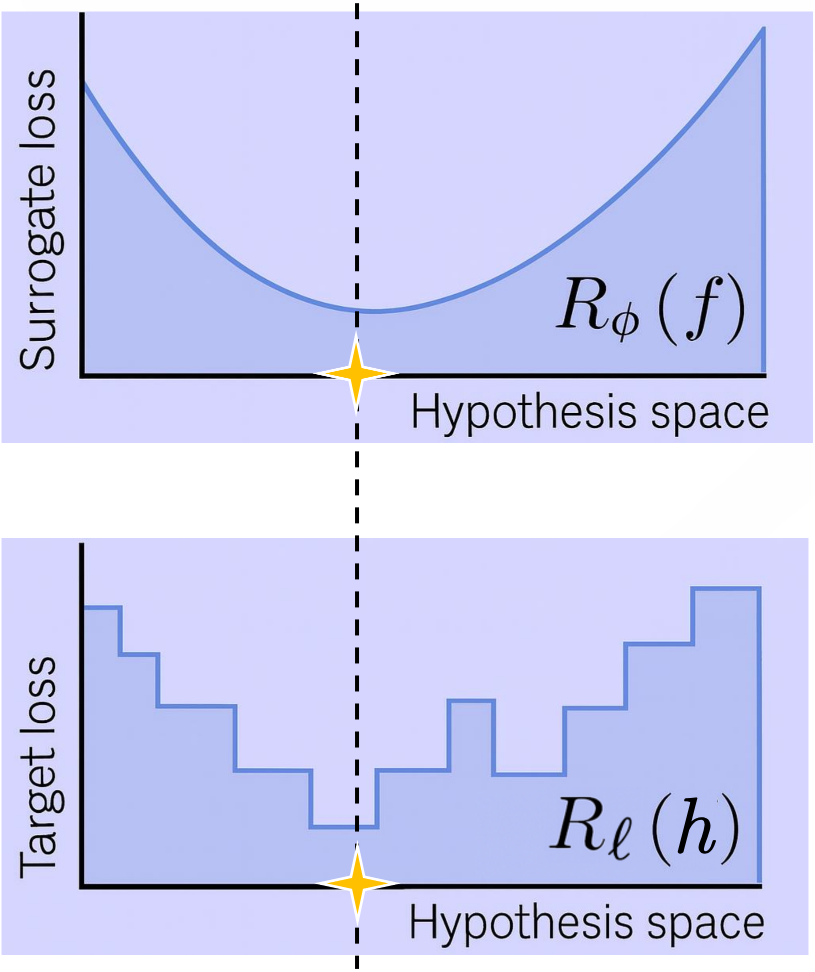
Continuous **Surrogate**

φ: $\mathbb{R}^d \to \widehat{\mathcal{Y}}$: prediction link. $h_\varphi := \varphi \circ f$: final predictive model.

What We Ask in Surrogate Loss

- Q1: Are surrogate and target losses minimized simultaneously? **Calibration/Fisher Consistency**

$$\underbrace{R_\phi(f) - \min_f R_\phi(f)}_{\text{(Surrogate) Regret}_\phi(f)} \to 0 \Rightarrow \underbrace{R_\ell(\varphi \circ f) - \min_h R_\ell(h)}_{\text{(Target) Regret}_\ell(\varphi \circ f)} \to 0$$



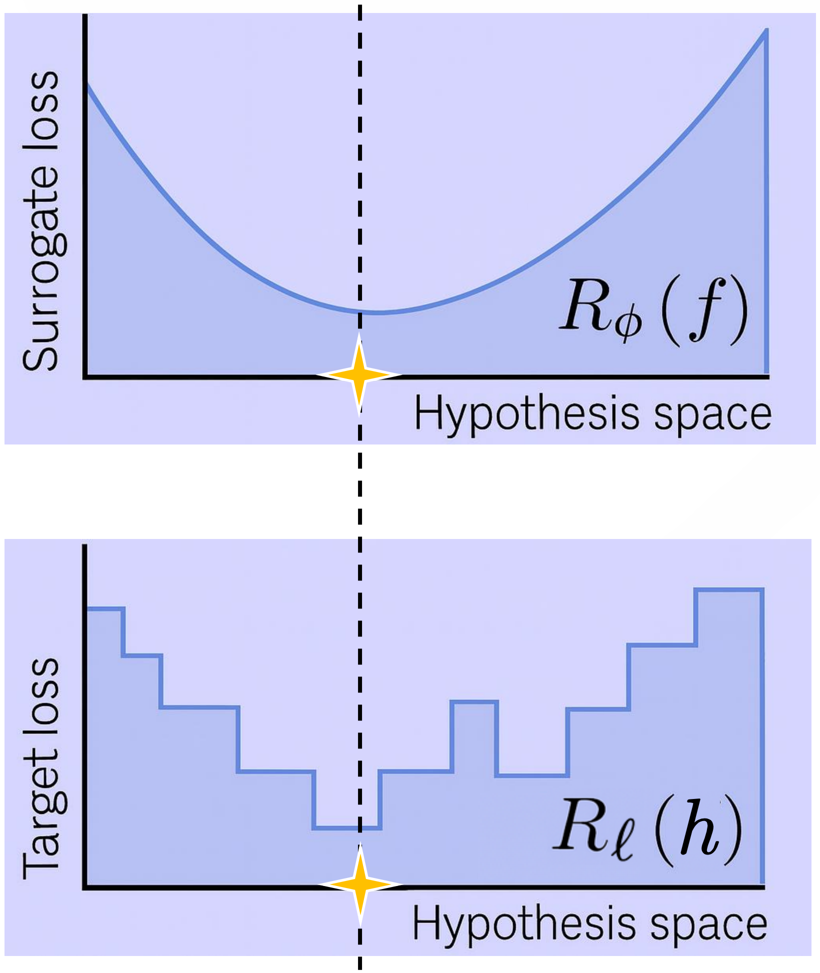
What We Ask in Surrogate Loss

- Q1: Are surrogate and target losses minimized simultaneously? **Calibration/Fisher Consistency**

$$\underbrace{R_\phi(f) - \min_f R_\phi(f)}_{\text{Regret}_\phi(f)} \to 0 \Rightarrow \underbrace{R_\ell(\varphi \circ f) - \min_h R_\ell(h)}_{\text{Regret}_\ell(\varphi \circ f)} \to 0$$

- Q2: How the convergence of $\text{Regret}_\phi(f)$ transfers to $\text{Regret}_\ell(\varphi \circ f)$? **Surrogate Regret Bound**

$$\text{Regret}_\ell(\varphi \circ f) \leq \psi(\text{Regret}_\phi(f))$$



What We Ask in Surrogate Loss

- Q2: How the convergence of $\mathrm{Regret}_{\phi}(f)$ transfers to $\mathrm{Regret}_{\ell}(\varphi \circ f)$?

$$\mathrm{Regret}_{\ell}(\varphi \circ f) \leq \psi(\mathrm{Regret}_{\phi}(f))$$

- Two typical bounds:
  - Square root:

    Deteriorates target regret convergence rate!

    $$\psi(*) = C \cdot \sqrt{*} \rightarrow \mathrm{Regret}_{\ell}(\varphi \circ f) = \mathcal{O}_p(1/n^{p/2})$$

  - Linear:

    Maintains fast target regret convergence rate!

    $$\psi(*) = C \cdot * \rightarrow \mathrm{Regret}_{\ell}(\varphi \circ f) = \mathcal{O}_p(1/n^{p})$$

What We Ask in Surrogate Loss

- Q2: How the convergence of $\mathrm{Regret}_\phi(f)$ transfers to $\mathrm{Regret}_\ell(\varphi \circ f)$?

$$\mathrm{Regret}_\ell(\varphi \circ f) \leq \psi(\mathrm{Regret}_\phi(f))$$

Linear ψ is desirable!

- Q3: Good Optimization Properties? **Convexity, Smoothness…**

A Negative Result

[FW21, Theorem 2] For surrogate-target loss pair (ϕ, ℓ) that is calibrated with surrogate regret bound ψ, if ϕ is **locally strongly convex and locally smooth**, the surrogate regret bound is **at least square-root**, e.g., there exists $\epsilon_0,\ C > 0$ that:

$$\psi(\epsilon) \geq C\sqrt{\epsilon},\ \forall \epsilon \leq \epsilon_0$$

Includes most existing convex smooth losses:

Cross-entropy/Focal/MSE/Binary Cross-Entropy/Dice/Jaccard Index…..

A popular conjecture: convexity, smoothness, and linear surrogate regret bound are **incompatible.**

Frongillo, R. and Waggoner, B,. (2021)
Surrogate regret bounds for polyhedral losses. Advances in Neural Information Processing Systems, 34:21569–21580,2021

A Negative Result

A popular conjecture: convexity, smoothness, and linear surrogate regret bound are **incompatible.**

This conjecture is overturned by **Convolutional Fenchel-Young Loss!**

- Works for **any** discrete target loss.
- Achieves a **linear surrogate regret bound**.
- **Smooth and convex**.
- Produces consistent **probability estimators**.

Target Loss Decomposition

- **Recap:** Target loss $\ell: \underset{\text{(Prediction space)}}{\widehat{\mathcal{Y}}} \times \underset{\text{(Label space)}}{\mathcal{Y}} \to \mathbb{R}_{\geq 0}, \; |\widehat{\mathcal{Y}}| = N, \; |\mathcal{Y}| = K.$
- **New concept:**

$\boldsymbol{\rho} - \boldsymbol{\ell}^{\boldsymbol{\rho}}$ **Decomposition:** $\ell(t, y) = \langle \boldsymbol{\rho}(y), \boldsymbol{\ell}^{\boldsymbol{\rho}}(t) \rangle + c(y)$

- Label encoding: $\boldsymbol{\rho}(y) \colon \mathcal{Y} \to \mathbb{R}^d$
- Loss encoding: $\boldsymbol{\ell}^{\boldsymbol{\rho}}(t) \colon \widehat{\mathcal{Y}} \to \mathbb{R}^d$
- Scalar offset: $c(y) \colon \mathcal{Y} \to \mathbb{R}$

An (always holds) example:

$$\boldsymbol{\rho}(y) = \boldsymbol{e}_y \in \mathbb{R}^K, \; \boldsymbol{\ell}^{\boldsymbol{\rho}}(t) = [\ell(t, 1), \cdots, \ell(t, K)]^\top, \; c = 0.$$

Fenchel-Young Loss

- **Recap:** label encoding $\boldsymbol{\rho}(y)\colon \mathcal{Y} \to \mathbb{R}^d$

Let $\Omega\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a **negentropy** (convex function) with $\mathrm{conv}\{\boldsymbol{\rho}(y)\}_{y=1}^{K} \subseteq \mathrm{dom}(\Omega)$

A **Fenchel-Young loss** [BMN20] $\phi_\Omega\colon \mathrm{dom}(\Omega^*) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ generated by Ω is defined as:

$$\phi_\Omega(\boldsymbol{\theta}, y) = \Omega(\boldsymbol{\rho}(y)) + \Omega^*(\boldsymbol{\theta}) - \langle \boldsymbol{\theta}, \boldsymbol{\rho}(y) \rangle$$

$\Omega^*(\boldsymbol{\theta}) = \sup_{\boldsymbol{\rho} \in \mathbb{R}^d} \langle \boldsymbol{\theta}, \boldsymbol{\rho} \rangle - \Omega(\boldsymbol{p})$ is Fenchel conjugate.

- Quantifies discrepancy between label $\rho(y)$ and score θ via **Fenchel-Young Inequality.**

$$\Omega(\boldsymbol{\rho}(y)) + \Omega^*(\boldsymbol{\theta}) \geq \langle \boldsymbol{\theta}, \boldsymbol{\rho}(y) \rangle$$

- Always **convex**, and **smooth** with strongly convex negentropy.

Blondel, M., Martins, A., &Niculae, V.. (2020). Learning with Fenchel–Young losses. Journal of Machine Learning Research, 21(35):1–69.

Linear Surrogate Regret Bound?

- **Duality between strong convexity and smoothness[KST09]:**

  ϕ_Ω is strongly convex if Ω is smooth.

[FW21, Theorem 2] For surrogate-target loss pair (ϕ, ℓ) that is calibrated with surrogate regret bound ψ, if ϕ is **locally strongly convex and locally smooth**, the surrogate regret bound is **at least square-root**, e.g., there exists $\epsilon_0,\ C > 0$ that:

$$\psi(\epsilon) \geq C\sqrt{\epsilon},\ \forall \epsilon \leq \epsilon_0$$

Smooth negentropy should be avoided!

Kakade, S., M., Shalev-Shwartz, S., & Tewari, A. (2009) On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. Technical report

Convolutional Entropy

Ordinary Negentropy $\Omega(\boldsymbol{p})$: commonly strongly convex & smooth (Square/Shannon)

Convolutional Entropy

Ordinary Negentropy $\Omega(\boldsymbol{p})$: commonly strongly convex & smooth (Square/Shannon)

New Concept-**Task Entropy:** $T(\boldsymbol{p}) := -\min_{t \in \hat{\mathcal{Y}}} \langle \boldsymbol{p}, \boldsymbol{\ell}^{\rho}(t) \rangle$

- **Recap:**
  - $\ell(t, y) = \langle \boldsymbol{\rho}(y), \boldsymbol{\ell}^{\rho}(t) \rangle + c(y)$
  - $T(\boldsymbol{p})$: a shifted negative minimum pointwise target risk.

Non-smooth but not strongly convex!

Convolutional Entropy

Ordinary Negentropy $\Omega(\boldsymbol{p})$: commonly strongly convex & smooth (Square/Shannon)

Task Entropy: $T(\boldsymbol{p}) := -\min_{t \in \widehat{\mathcal{Y}}} \langle \boldsymbol{p}, \boldsymbol{\ell}^{\boldsymbol{\rho}}(t) \rangle$

Convolutional Entropy: $\Omega_T(\boldsymbol{p}) := (\Omega + T)(\boldsymbol{p})$ **strongly convex and non-smooth!**

Convolutional Entropy

Ordinary Negentropy $\Omega(\boldsymbol{p})$: commonly strongly convex & smooth (Square/Shannon)

Task Entropy: $T(\boldsymbol{p}) := -\min_{t \in \widehat{\mathcal{Y}}} \langle \boldsymbol{p}, \boldsymbol{\ell}^{\boldsymbol{\rho}}(t) \rangle$

Convolutional Entropy: $\Omega_T(\boldsymbol{p}) := (\Omega + T)(\boldsymbol{p})$ **strongly convex and non-smooth!**

Why "Convolutional"?

$$\Omega_T^*(\boldsymbol{\theta}) = (\Omega^* \square T^*)(\boldsymbol{\theta}) = \inf_{\boldsymbol{u}} \{\Omega^*(\boldsymbol{\theta} - \boldsymbol{u}) + T^*(\boldsymbol{u})\}$$

Infimal Convolution!

Convolutional Entropy

Ordinary Negentropy $\Omega(\boldsymbol{p})$: commonly strongly convex & smooth (Square/Shannon)

Task Entropy: $T(\boldsymbol{p}) := -\min_{t \in \widehat{\mathcal{Y}}} \langle \boldsymbol{p}, \boldsymbol{\ell}^{\rho}(t) \rangle$

Convolutional Entropy: $\Omega_T(\boldsymbol{p}) := (\Omega + T)(\boldsymbol{p})$ **strongly convex and non-smooth!**

$$\Omega_T^*(\boldsymbol{\theta}) = (\Omega^* \square T^*)(\boldsymbol{\theta}) = \inf_{\boldsymbol{u}} \{\Omega^*(\boldsymbol{\theta} - \boldsymbol{u}) + T^*(\boldsymbol{u})\} = \inf_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\rho} \boldsymbol{\pi})$$

$$\mathcal{L}^{\rho} := [\boldsymbol{\ell}^{\rho}(1), \cdots, \boldsymbol{\ell}^{\rho}(N)]^{\top}$$

Convolutional Entropy

Ordinary Negentropy $\Omega(\boldsymbol{p})$: commonly strongly convex & smooth (Square/Shannon)

Task Entropy: $T(\boldsymbol{p}) := -\min_{t \in \widehat{\mathcal{Y}}} \langle \boldsymbol{p}, \boldsymbol{\ell}^{\rho}(t) \rangle$

Convolutional Entropy: $\Omega_T(\boldsymbol{p}) := (\Omega + T)(\boldsymbol{p})$ **strongly convex and non-smooth!**

Conjugated Convolutional Entropy: $\Omega_T^*(\boldsymbol{\theta}) = \inf_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\rho} \boldsymbol{\pi})$

Convolutional Fenchel-Young Loss

For strongly convex $\Omega\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with $\mathrm{conv}\{\boldsymbol{\rho}(y)\}_{y=1}^{K} \subseteq \mathrm{dom}(\Omega)$ and $\mathrm{dom}(\Omega^*) = \mathbb{R}^d$, a **Convolutional Fenchel-Young loss** $\phi_{\Omega_T}\colon \mathrm{dom}(\Omega_T^*) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ generated by Ω_T is:

$$\phi_{\Omega_T}(\boldsymbol{\theta}, y) = \Omega_T(\boldsymbol{\rho}(y)) + \inf_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\rho}\boldsymbol{\pi}) - \langle \boldsymbol{\theta}, \boldsymbol{\rho}(y) \rangle$$

Convolutional Fenchel-Young Loss

For strongly convex $\Omega: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with $\mathrm{conv}\{\boldsymbol{\rho}(y)\}_{y=1}^K \subseteq \mathrm{dom}(\Omega)$ and $\mathrm{dom}(\Omega^*) = \mathbb{R}^d$, a **Convolutional Fenchel-Young loss** ϕ_{Ω_T}: $\mathrm{dom}(\Omega_T^*) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ generated by Ω_T is:

$$\phi_{\Omega_T}(\boldsymbol{\theta}, y) = \Omega_T(\boldsymbol{\rho}(y)) + \underbrace{\inf_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\rho} \boldsymbol{\pi})}_{\text{Attainable and efficiently solvable (Lemma 8)}} - \langle \boldsymbol{\theta}, \boldsymbol{\rho}(y) \rangle$$

Attainable and efficiently solvable (Lemma 8)

Convolutional Fenchel-Young Loss

For strongly convex $\Omega: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with $\text{conv}\{\boldsymbol{\rho}(y)\}_{y=1}^K \subseteq \text{dom}(\Omega)$ and $\text{dom}(\Omega^*) = \mathbb{R}^d$, a **Convolutional Fenchel-Young loss** ϕ_{Ω_T}: $\text{dom}(\Omega_T^*) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ generated by Ω_T is:

$$\phi_{\Omega_T}(\boldsymbol{\theta}, y) = \Omega_T(\boldsymbol{\rho}(y)) + \min_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\rho} \boldsymbol{\pi}) - \langle \boldsymbol{\theta}, \boldsymbol{\rho}(y) \rangle$$

Convolutional Fenchel-Young Loss

For strongly convex $\Omega: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with $\mathrm{conv}\{\boldsymbol{\rho}(y)\}_{y=1}^{K} \subseteq \mathrm{dom}(\Omega)$ and $\mathrm{dom}(\Omega^*) = \mathbb{R}^d$, a **Convolutional Fenchel-Young loss** ϕ_{Ω_T}: $\mathrm{dom}(\Omega_T^*) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ generated by Ω_T is:

$$\phi_{\Omega_T}(\boldsymbol{\theta}, y) = \Omega_T(\boldsymbol{\rho}(y)) + \min_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) - \langle \boldsymbol{\theta}, \boldsymbol{\rho}(y) \rangle$$

- (Full domain) $\mathrm{dom}(\phi_{\Omega_T}) = \mathbb{R}^d$.
- **(Smoothness and convexity)** ϕ_{Ω_T} is convex and smooth.
- **(Envelope theorem)**
  $$\nabla_{\boldsymbol{\theta}} \min_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) = \nabla \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}^*), \ \forall \boldsymbol{\pi}^* \in \Pi(\boldsymbol{\theta}) := \operatorname*{argmin}_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}^*)$$

Convolutional Fenchel-Young Loss

For strongly convex $\Omega: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with $\mathrm{conv}\{\boldsymbol{\rho}(y)\}_{y=1}^K \subseteq \mathrm{dom}(\Omega)$ and $\mathrm{dom}(\Omega^*) = \mathbb{R}^d$, a **Convolutional Fenchel-Young loss** $\phi_{\Omega_T}: \mathrm{dom}(\Omega_T^*) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ generated by Ω_T is:

$$\phi_{\Omega_T}(\boldsymbol{\theta}, y) = \Omega_T(\boldsymbol{\rho}(y)) + \min_{\boldsymbol{\pi} \in \Delta^N} \Omega^*(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}) - \langle \boldsymbol{\theta}, \boldsymbol{\rho}(y) \rangle$$

- **Probability estimator:** For any $\boldsymbol{\eta} \in \mathrm{relint}(\Delta^K)$, the pointwise surrogate risk $R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) := \mathbb{E}_{Y \sim \boldsymbol{\eta}}[\phi_{\Omega_T}(\boldsymbol{\theta}, Y)]$ is minimized at $\boldsymbol{\theta}^*$:

$$\mathbb{E}_{Y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(Y)] = \nabla \Omega^*(\boldsymbol{\theta}^* + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}^*), \ \forall \boldsymbol{\pi}^* \in \Pi(\boldsymbol{\theta}^*)$$

(e.g., $\mathbb{E}_{Y \sim \boldsymbol{\eta}}[\boldsymbol{\rho}(Y)] = \boldsymbol{\eta}$ **when** $\boldsymbol{\rho}(y) = \boldsymbol{e}_y \in \mathbb{R}^K$**)**

Surrogate Regret Analysis

- **Recap:** Surrogate regret: $\mathrm{Regret}_{\phi_{\Omega_T}}(h) := R_{\phi_{\Omega_T}}(h) - \min_h R_{\phi_{\Omega_T}}(h)$

$$= \mathbb{E}_X[\underbrace{R_{\phi_{\Omega_T}}(h(X), \boldsymbol{\eta}(X)) - \min_{\boldsymbol{\theta} \in \mathbb{R}^d} R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}(X))}_{\mathrm{Regret}(h(X), \boldsymbol{\eta}(X)):\ \text{Pointwise Surrogate Regret}}]$$

- **Pointwise surrogate regret of Conv-FY loss:**

$$\mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \min_{\boldsymbol{\theta} \in \mathbb{R}^d} R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta})$$

Surrogate Regret Analysis

- **Recap:** Surrogate regret: $\mathrm{Regret}_{\phi_{\Omega_T}}(h) := R_{\phi_{\Omega_T}}(h) - \min_h R_{\phi_{\Omega_T}}(h)$

$$= \mathbb{E}_X\Big[\underbrace{R_{\phi_{\Omega_T}}(h(X), \boldsymbol{\eta}(X)) - \min_{\boldsymbol{\theta} \in \mathbb{R}^d} R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}(X))}_{\mathrm{Regret}(h(X), \boldsymbol{\eta}(X)):\ \text{Pointwise Surrogate Regret}}\Big]$$

- **Pointwise surrogate regret of Conv-FY loss:**

$$\mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \min_{\boldsymbol{\theta} \in \mathbb{R}^d} R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta})$$

Surrogate regret decomposition:

$$\mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = R_{\phi_\Omega}(\boldsymbol{\theta} + \mathcal{L}^{\boldsymbol{\rho}} \boldsymbol{\pi}^*, \boldsymbol{\eta}) + \sum_{t=1}^{N} \pi^*_t \mathrm{Regret}_\ell(t, \boldsymbol{\eta}), \quad \forall \boldsymbol{\pi}^* \in \Pi(\boldsymbol{\theta})$$

Surrogate Regret Analysis

- **Recap:** Surrogate regret: $\mathrm{Regret}_{\phi_{\Omega_T}}(h) := R_{\phi_{\Omega_T}}(h) - \min_h R_{\phi_{\Omega_T}}(h)$

$$= \mathbb{E}_X[\underbrace{R_{\phi_{\Omega_T}}(h(X), \boldsymbol{\eta}(X)) - \min_{\boldsymbol{\theta} \in \mathbb{R}^d} R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}(X))}]$$

$\mathrm{Regret}(h(X), \boldsymbol{\eta}(X))$: Pointwise Surrogate Regret

- **Pointwise surrogate regret of Conv-FY loss:**

$$\mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \min_{\boldsymbol{\theta} \in \mathbb{R}^d} R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta})$$

Surrogate regret decomposition:

$$\mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \underbrace{R_{\phi_\Omega}(\boldsymbol{\theta} + \mathcal{L}^\rho \boldsymbol{\pi}^*, \boldsymbol{\eta})}_{\text{Risk of FY loss } \phi_\Omega \ (\geq 0)} + \underbrace{\sum_{t=1}^{N} \pi^*_t \mathrm{Regret}_\ell(t, \boldsymbol{\eta})}_{\text{Convex combination of target regret}}, \quad \forall \boldsymbol{\pi}^* \in \Pi(\boldsymbol{\theta})$$

Surrogate Regret Analysis

- **Recap:** Surrogate regret: $\mathrm{Regret}_{\phi_{\Omega_T}}(h) := R_{\phi_{\Omega_T}}(h) - \min_h R_{\phi_{\Omega_T}}(h)$

$$= \mathbb{E}_X\Big[\underbrace{R_{\phi_{\Omega_T}}(h(X), \boldsymbol{\eta}(X)) - \min_{\boldsymbol{\theta} \in \mathbb{R}^d} R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}(X))}_{\mathrm{Regret}(h(X), \boldsymbol{\eta}(X)):\ \text{Pointwise Surrogate Regret}}\Big]$$

- **Pointwise surrogate regret of Conv-FY loss:**

$$\mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \min_{\boldsymbol{\theta} \in \mathbb{R}^d} R_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta})$$

Surrogate regret decomposition:

$$\mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \geq R_{\phi_\Omega}(\boldsymbol{\theta} + \mathcal{L}^\rho \boldsymbol{\pi}^*, \boldsymbol{\eta}) + \sum_{t=1}^{N} \pi^*_t \mathrm{Regret}_\ell(t, \boldsymbol{\eta}), \quad \forall \boldsymbol{\pi}^* \in \Pi(\boldsymbol{\theta})$$

Surrogate Regret Analysis

$$\text{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \geq \sum_{t=1}^{N} \pi^*(t)\, \text{Regret}_{\ell}(t, \boldsymbol{\eta})$$

$$\forall \hat{t} \in \underset{t \in \hat{\mathcal{Y}}}{\text{argmax}}\, \pi^*(t) \colon\ \pi^*(\hat{t}) \geq 1/N \text{ since } \boldsymbol{\pi}^* \in \Delta^N$$

Surrogate Regret Analysis

$$\text{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \geq \sum_{t=1}^{N} \pi^*(t)\, \text{Regret}_\ell(t, \boldsymbol{\eta})$$

$$\Downarrow$$

$$\forall \hat{t} \in \operatorname*{argmax}_{t \in \hat{\mathcal{Y}}} \pi^*(t):\ \pi^*(\hat{t}) \geq 1/N \text{ since } \boldsymbol{\pi}^* \in \Delta^N$$

$$\text{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \geq \text{Regret}_\ell\left(\hat{t}, \boldsymbol{\eta}\right)/N$$

Linear Regret Link Construction

$$\mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta},\boldsymbol{\eta}) \geq \sum_{t=1}^{N} \pi^*(t)\,\mathrm{Regret}_{\ell}(t,\boldsymbol{\eta})$$

$$\Downarrow$$

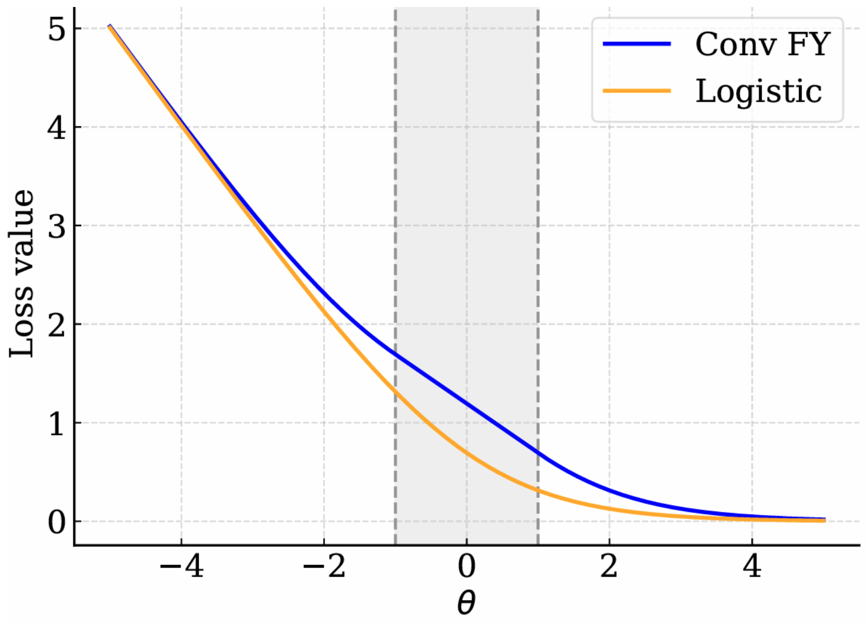
$\forall \hat{t} \in \operatorname{argmax}_{t\in\hat{\mathcal{Y}}} \pi^*(t)$: $\pi^*(\hat{t}) \geq 1/N$ since $\boldsymbol{\pi}^* \in \Delta^N$

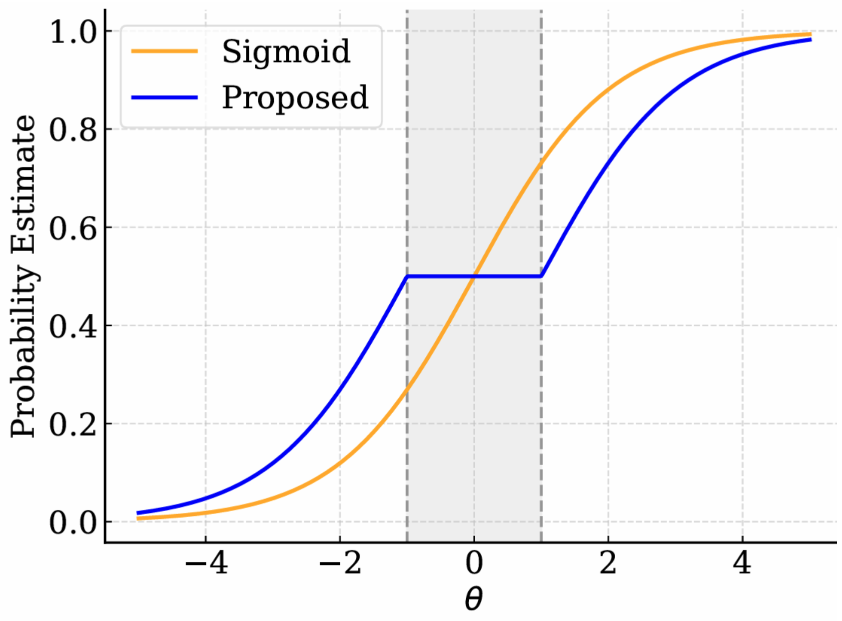
$$\mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta},\boldsymbol{\eta}) \geq \mathrm{Regret}_{\ell}\left(\hat{t},\boldsymbol{\eta}\right)/N$$

($\boldsymbol{\pi}$-argmax link): For function $\boldsymbol{\pi}^*: \mathbb{R}^d \to \Delta^N$, if $\boldsymbol{\pi}^*(\boldsymbol{\theta}) \in \Pi(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \mathbb{R}^d$:

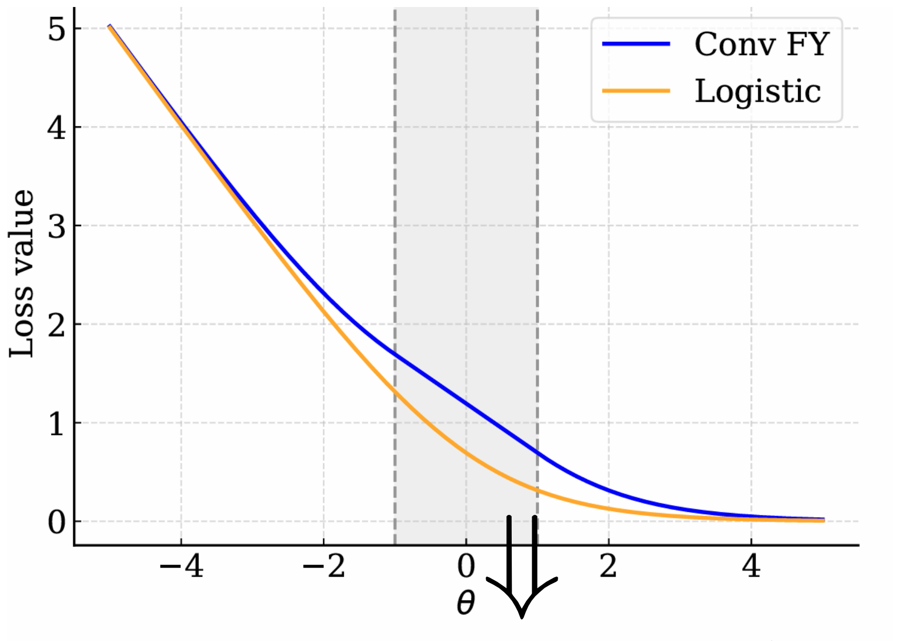
$$\mathrm{Regret}_{\ell}(\varphi(\boldsymbol{\theta}),\boldsymbol{\eta}) \leq N\ \mathrm{Regret}_{\phi_{\Omega_T}}(\boldsymbol{\theta},\boldsymbol{\eta}), \quad \forall(\boldsymbol{\theta},\boldsymbol{\eta}) \in \mathbb{R}^d \times \Delta^K, \quad \varphi := \operatorname{argmax} \circ \boldsymbol{\pi}^*$$

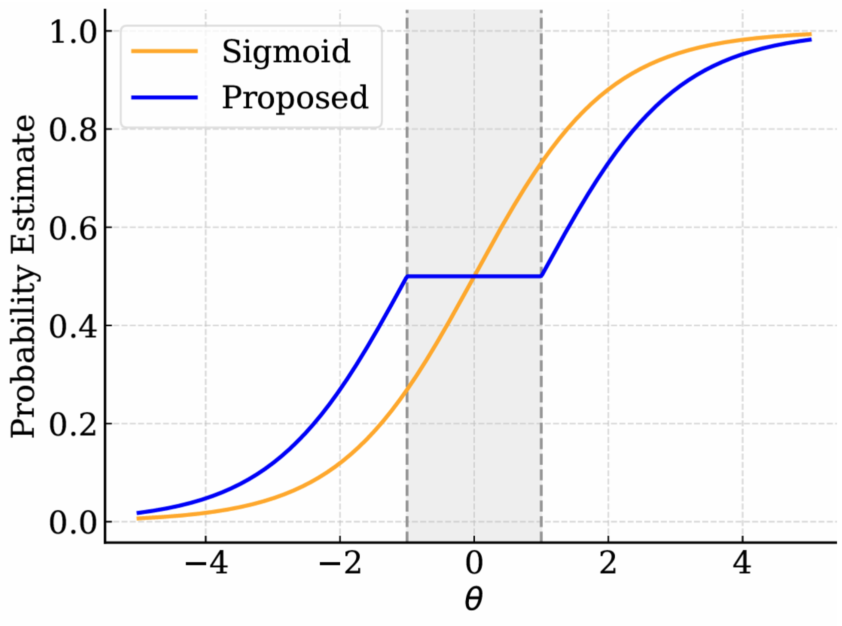
Binary Case Visualization





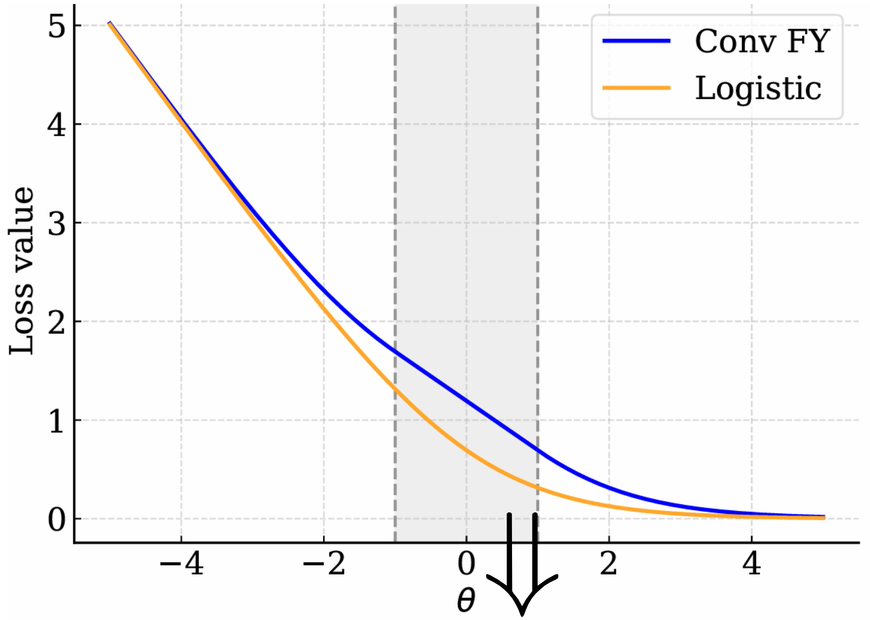
Binary Case Visualization

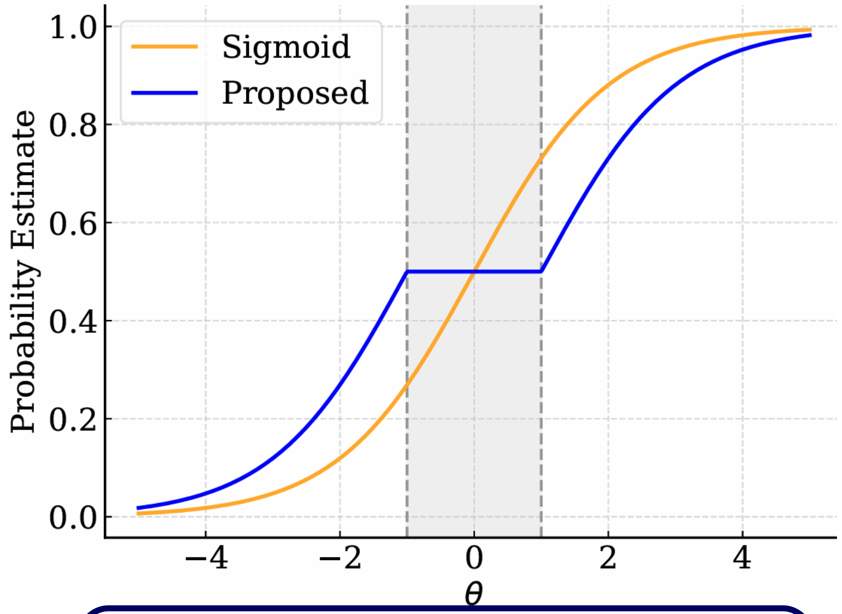




Linear here!

Binary Case Visualization





Linear here!

Circumventing sqrt bound by **injecting linearity!**

[FW21, Theorem 2] ... if ϕ is **locally strongly convex and locally smooth**, the surrogate regret bound is **at least square-root…….**

Frongillo, R. and Waggoner, B,. (2021)
Surrogate regret bounds for polyhedral losses. Advances in Neural Information Processing Systems, 34:21569–21580,2021