Surrogate Regret Bounds for Polyhedral Losses

Rafael Frongillo and **Bo Waggoner** University of Colorado, Boulder **NeurIPS 2021**

The short version

Surrogate risk minimization for supervised learning:

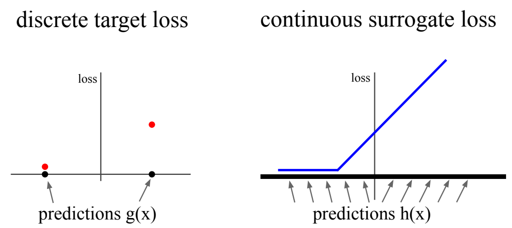
- Goal: optimize **discrete "target"** loss $\ell(r, y)$ *classification, structured prediction*
  r = prediction, y = label
- Approach: optimize **continuous "surrogate"** loss $L(u, y)$
  then "link" to target prediction $r = \psi(u)$

Q: When does (quick) **surrogate** convergence imply (quick) **target** convergence?
"Surrogate regret bounds" or "regret transfer rates"

$$\mathrm{Regret}_\ell(\psi \circ h) \le \zeta\Big(\mathrm{Regret}_L(h)\Big)$$

Regret = "excess risk" over Bayes optimal

This paper: **polyhedral surrogates** *piecewise-linear and convex*



Theorem 1: All polyhedral surrogates have **linear** regret transfer rates!
Polyhedral: **surrogate regret** $\le \epsilon \implies$ **target regret** $\le O(\epsilon)$.

Theorem 2: Sufficiently "non-polyhedral" transfers are **quadratically slower**.
Non-polyhedral: Can have **surrogate regret** $\le \epsilon$ yet **target regret** $\ge \Omega(\sqrt{\epsilon})$.

Background: surrogate risk, polyhedral losses

Data: $(x, y) \in \mathcal{X} \times \mathcal{Y}$ *from distributions $\mathcal{D}$*

Hypotheses: $g : \mathcal{X} \to \mathcal{R}$ *example soon*

Discrete target loss: $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_{\ge 0}$. *discrete: $\mathcal{R}$ is finite*
Classification ($0-1$ loss), ranking, top-k

Regret: $\mathrm{Regret}_\ell(g; \mathcal{D}) := \mathbb{E}_{x,y\sim\mathcal{D}}\left[\ell(g(x), y) - \ell(g^*(x), y)\right]$. *$g^*$ = Bayes optimal*

Continuous surrogate loss: $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_{\ge 0}$. *for some d*
Hinge loss for classification, BEP surrogate (Ramaswamy et al. 2018)

Polyhedral surrogate $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_{\ge 0}$:
pointwise maximum of a finite set of affine functions.

Recent work: **polyhedral surrogates** are natural convexifications of **discrete losses**.

- Finocchiaro, Frongillo, Waggoner 2019, 2020; applications e.g. Wang, Scott 2020.

Desirable surrogates:

- Optimizable. *Convex, strongly convex, etc.*
- Generalization bound, convergence rate.
- **Connection to target...**

Regret transfer function $\zeta : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$: *Continuous at 0, $\zeta(0) = 0$*

$$\mathrm{Regret}_\ell(\psi \circ h; \mathcal{D}) \;\le\; \zeta\Big(\mathrm{Regret}_L(h; \mathcal{D})\Big).$$

If **any such** ζ exists, (L, ψ) is **consistent** for ℓ.
As $\mathrm{Regret}_L(h^{(n)}; \mathcal{D}) \to 0$, we have $\mathrm{Regret}_\ell^{(n)}(\psi \circ h^{(n)}; \mathcal{D}) \to 0$.

Ideally: **fast** convergence, e.g. $\zeta(\epsilon) = O(\sqrt{\epsilon})$ (good), $O(\epsilon^{3/4})$ (better), $O(\epsilon)$ (best).

Prior work: binary classification (Zhang et al. 2004, etc.); bipartite ranking (Agarwal 2014, etc.); multiclass classification (Duchi et al. 2018, etc.); hierarchical classification (Ramaswamy et al. 2015); strongly convex surrogates (Nowak-Vila et al. 2019, etc.).

Results

Theorem 1. Suppose L is polyhedral and (L, ψ) are consistent for the discrete target ℓ. Then there exists $C > 0$ such that, for all $\mathcal{D}$ and h,

$$\mathrm{Regret}_\ell(\psi \circ h; \mathcal{D}) \le C \cdot \mathrm{Regret}_L(h; \mathcal{D}).$$

Proof sketch. Fix x; let $p, q \in \Delta_\mathcal{Y}$ be distributions of y given x.

- Suffices to prove that for all p and u, $\mathrm{Regret}_\ell(\psi(u); p) \le C \cdot \mathrm{Regret}_L(u; p)$.
- $\forall q$, $\exists \alpha_q > 0$ such that $\mathrm{Regret}_\ell(\psi(u); q) \le \alpha_q \cdot \mathrm{Regret}_L(u; q)$.
- Polyhedral losses have a **finite structure**: *related to embeddings framework*
  - There is a finite $U \subseteq \mathbb{R}^d$ such that, for all p, U contains an optimal prediction.
  - These U partition $\Delta_\mathcal{Y}$ into finitely many polytopes.
- Regret is linear on each polytope.
- Therefore $\mathrm{Regret}_L(u; p)$ is a convex combination over the corners.
- So we can take $C = \max_{\text{corners } q} \alpha_q$.

Theorem 2. Let L be a locally strongly convex surrogate with locally Lipschitz gradient. Suppose (L, ψ) is consistent for ℓ with:

$$\mathrm{Regret}_\ell(\psi \circ h) \le \zeta\Big(\mathrm{Regret}_L(h)\Big).$$

Then there exists $c > 0$ such that, for all small enough $\epsilon > 0$,

$$\zeta(\epsilon) \ge c\sqrt{\epsilon}.$$

Proof idea:

- Fix a "boundary" prediction u_0.
- Consider a conditional distributions $\{p_\lambda : 0 \le \lambda \le 1\}$
- **Target regret** shrinks linearly as $\lambda \to 0$, but **surrogate regret** shrinks with $\sqrt{\lambda}$.

Examples: exponential loss, Huber loss *via strengthenings*

Investigating the constant in Theorem 1

$C = \max_{\text{corners } q} \alpha_q$.

What is C?

Can bound $C \le \beta_L \cdot \beta_\ell \cdot \beta_\psi$.

- β_L comes from **Hoffman constants**. *minimum slope of polyhedral losses*
- β_ℓ comes from a simple maximum possible regret.
- $\beta_\psi = \frac{1}{\epsilon}$ where the link is ϵ-separated. *see embedding framework*

Big picture

Polyhedral surrogate losses are nice:

- Consistent polyhedral surrogates always exist
- Embedding framework
- **(This work) Always satisfy linear regret transfer** *a.k.a. calibration functions*

Next questions:

- Are they good for the whole pipeline? (optimization + generalization)
- Applications...
- e.g. low-dimensional or otherwise "nice" polyhedral surrogates