Algorithmic Foundations of Learning

Lecture 8
Convex Loss Surrogates. Elements of Convex Theory

Patrick Rebeschini

Department of Statistics
University of Oxford

Recall Results on Binary Classification

- $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$
- Admissible action set $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \to \{-1, 1\}\}$
- **True** loss function $\ell(a, (x, y)) = 1_{a(x) \neq y} = \varphi^\star(a(x)y)$ with $\varphi^\star(u) := 1_{u \leq 0}$

$$r(a) = \mathbf{P}(a(X) \neq Y) \qquad a^\star \in \operatorname*{argmin}_{a \in \mathcal{A}} r(a) \qquad a^{\star\star} \in \operatorname*{argmin}_{a \in \mathcal{B}} r(a)$$

$$R(a) = \frac{1}{n} \sum_{i=1}^{n} 1_{a(X_i) \neq Y_i} \qquad A^\star \in \operatorname*{argmin}_{a \in \mathcal{A}} R(a)$$

So far we have proved:

$$\boxed{\mathbf{P}\left(r(A^\star) - r(a^\star) \lesssim \sqrt{\frac{\mathtt{VC}(\mathcal{A})}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) \geq 1 - \delta}$$

Problem: In general, computing $A^\star$ is NP hard!

Idea: Define convex relaxation of the original problem

Convexity

Convex function (Definition 8.1)

A function $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* if for every $x, \tilde{x} \in \mathbb{R}^d, \lambda \in [0, 1]$ we have

$$\boxed{f(\lambda x + (1 - \lambda)\tilde{x}) \leq \lambda f(x) + (1 - \lambda) f(\tilde{x})}$$

Convex set (Definition 8.2)

A set $\mathcal{A}$ is *convex* if for every $a, \tilde{a} \in \mathcal{A}, \lambda \in [0, 1]$ we have

$$\boxed{\lambda a + (1 - \lambda)\tilde{a} \in \mathcal{A}}$$

Convex Loss Surrogates

Convex loss surrogate (Definition 8.3)

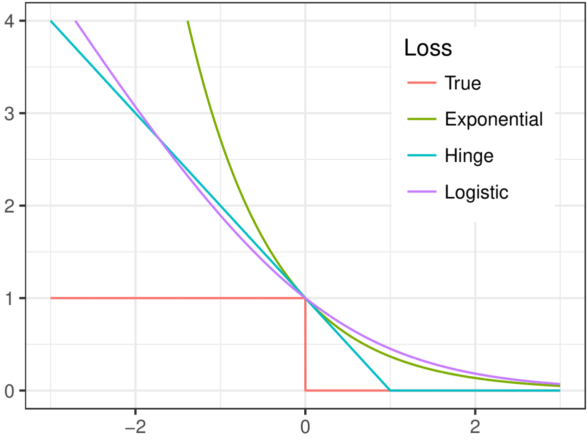
A function $\varphi : \mathbb{R} \to \mathbb{R}_+$ is called a *convex loss surrogate* if:
- convex
- non-increasing
- $\varphi(0) = 1$

True loss:
$\varphi^\star(u) = 1_{u \leq 0}$

Exponential loss:
$\varphi(u) = e^{-u}$

Hinge loss:
$\varphi(u) = \max\{1 - u, 0\}$

Logistic loss:
$\varphi(u) = \log_2(1 + e^{-u})$



Convex Soft Classifiers

- **Soft** classifiers $\mathcal{A}_{\text{soft}} \subseteq \mathcal{B}_{\text{soft}} := \{a : \mathbb{R}^d \to \mathbb{R}\}$
- If $a \in \mathcal{B}_{\text{soft}}$, corresponding **hard** classifier is given by $\operatorname{sign}(a)$

1. **Linear functions with convex parameter space:**

$$\mathcal{A}_{\text{soft}} = \{a(x) = w^\top x + b : w \in \mathcal{C}_1 \subseteq \mathbb{R}^d, b \in \mathcal{C}_2 \subseteq \mathbb{R}\}$$

   $\mathcal{C}_1, \mathcal{C}_2$ are convex sets

2. **Majority votes (Boosting):**

$$\mathcal{A}_{\text{soft}} = \{a(x) = \sum_{i=1}^{m} w_j h_j(x) : w = (w_1, \ldots, w_m) \in \Delta_m\}$$

   Δ_m is the m-dim. simplex and $h_1, \ldots, h_m : \mathbb{R}^d \to \mathbb{R}$ are *base classifiers*

Empirical φ-Risk Minimization

If φ and $\mathcal{A}_{\text{soft}}$ are convex, we are left with a convex problem

$$R_\varphi(a) = \frac{1}{n} \sum_{i=1}^{n} \varphi(a(X_i) Y_i)$$

$$A^\star_\varphi \in \operatorname*{argmin}_{a \in \mathcal{A}_{\text{soft}}} R_\varphi(a)$$

Zhang's Lemma

$$r_\varphi(a) = \mathbf{E}\,\varphi(a(X)Y) \qquad a^{\star\star}_\varphi \in \operatorname*{argmin}_{a\in\mathcal{B}_{\text{soft}}} r_\varphi(a)$$

$$r(a) = \mathbf{E}\,\varphi^\star(a(X)Y) = \mathbf{P}(a(X) \neq Y) \qquad a^{\star\star} \in \operatorname*{argmin}_{a\in\mathcal{B}} r(a)$$

Zhang's Lemma (Lemma 8.5)

Let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be a convex loss surrogate. For any $\tilde\eta \in [0,1]$, $\tilde a \in \mathbb{R}$, let

$$H_{\tilde\eta}(\tilde a) := \varphi(\tilde a)\tilde\eta + \varphi(-\tilde a)(1-\tilde\eta), \qquad \tau(\tilde\eta) := \inf_{\tilde a\in\mathbb{R}} H_{\tilde\eta}(\tilde\alpha).$$

Assume that there exist $c > 0$ and $\nu \in [0,1]$ such that

$$\boxed{\left|\tilde\eta - \frac{1}{2}\right| \le c(1-\tau(\tilde\eta))^\nu \qquad \text{for any } \tilde\eta \in [0,1]}$$

Then, for any $a : \mathbb{R}^d \to \mathbb{R}$ we have

$$\boxed{\underbrace{r(\operatorname{sign}(a)) - r(a^{\star\star})}_{\substack{\text{excess risk}\\ \text{hard classifier}}} \le 2c(\underbrace{r_\varphi(a) - r_\varphi(a^{\star\star}_\varphi)}_{\substack{\text{excess } \varphi\text{-risk}\\ \text{soft classifier}}})^\nu}$$

Zhang's Lemma: Examples

- **Exponential loss:**
  $\tau(\tilde{\eta}) = 2\sqrt{\tilde{\eta}(1-\tilde{\eta})}$
  $c = 1/\sqrt{2}$
  $\nu = 1/2$

- **Hinge loss:**
  $\tau(\tilde{\eta}) = 1 - |1 - 2\tilde{\eta}|$
  $c = 1/2$
  $\nu = 1$

- **Logistic loss:**
  $\tau(\tilde{\eta}) = -\tilde{\eta}\log_2 \tilde{\eta} - (1-\tilde{\eta})\log_2(1-\tilde{\eta})$
  $c = 1/\sqrt{2}$
  $\nu = 1/2$

Zhang's Lemma shows that we can reliably focus on convex problems

Elements of Convex Theory

Subgradients (Definition 8.8)

Let $f : \mathcal{C} \subset \mathbb{R}^d \to \mathbb{R}$. A vector $g \in \mathbb{R}^d$ is a *subgradient* of f at $x \in \mathcal{C}$ if

$$\boxed{f(x) - f(y) \leq g^T(x - y) \qquad \text{for any } y \in \mathcal{C}}$$

The set of subgradients of f at x is denoted $\partial f(x)$.

Subgradients yield **global** information (**uniform** lower bounds)

Convexity and subgradients (Theorem 8.9)

Let $f : \mathcal{C} \subseteq \mathbb{R}^d \to \mathbb{R}$ with $\mathcal{C}$ convex:

f is convex $\Longrightarrow$ for any $x \in \text{int}(\mathcal{C}), \partial f(x) \neq \emptyset$

f is convex $\Longleftarrow$ for any $x \in \mathcal{C}, \partial f(x) \neq \emptyset$

If f is convex and differentiable at x, then $\nabla f(x) \in \partial f(x)$

Convex functions that are differentiable allow to infer **global** information (i.e., subgradients) from **local** information (i.e., gradients)

This is why convex problems are “typically” amenable to computations...
To prove algorithms converge we need additional local-to-global properties

Are Convex Problems Easy to Solve?

- *Convex hull*: $\mathrm{conv}(\mathcal{T}) := \left\{ \sum_{j=1}^m w_j t_j : w \in \Delta_m, t_1, \dots, t_m \in \mathcal{T}, m \in \mathbb{N} \right\}$
- *Epigraph*: $\mathrm{epi}(f) := \{(x,t) \in \mathcal{D} \times \mathbb{R} : f(x) \le t\}$.

Proposition 8.6

$$\min_{t \in \mathcal{T}} c^\top t = \min_{t \in \mathrm{conv}(\mathcal{T})} c^\top t, \qquad \max_{t \in \mathcal{T}} c^\top t = \max_{t \in \mathrm{conv}(\mathcal{T})} c^\top t.$$

Proof: As $\mathcal{T} \subseteq \mathrm{conv}(\mathcal{T})$, we have $\min_{t \in \mathcal{T}} c^\top t \ge \min_{t \in \mathrm{conv}(\mathcal{T})} c^\top t$. Other direction:

$$\begin{aligned}
\min_{t \in \mathrm{conv}(\mathcal{T})} c^\top t &= \min_{m \in \mathbb{N}, t_1, \dots, t_m \in \mathcal{T}, (w_1, \dots, w_m) \in \Delta_m} c^\top \left(\sum_{j=1}^m w_j t_j \right) \\
&= \min_{m \in \mathbb{N}, t_1, \dots, t_m \in \mathcal{T}, (w_1, \dots, w_m) \in \Delta_m} \sum_{j=1}^m w_j c^\top t_j \ge \min_{t \in \mathcal{T}} c^\top t.
\end{aligned}$$

Proposition 8.7

For any $f : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}$, $\min_{x \in \mathcal{D}} f(x) = \min_{(x,t) \in \mathcal{C}} t$ with $\mathcal{C} = \mathrm{conv}(\mathrm{epi}(f))$.

Any minimization problem can be written in a convex form!

Local-to-Global Properties

- **Convex:** $f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \mathbb{R}^d$
- α**-Strongly Convex:**

  $\exists \alpha > 0$ such that $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2} \|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^d$
- β**-Smooth:**

  $\exists \beta > 0$ such that $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^d$
- γ**-Lipschitz:**

  $\exists \gamma > 0$ such that $f(x) - \gamma \|y - x\|_2 \leq f(y) \leq f(x) + \gamma \|y - x\|_2 \; \forall x, y \in \mathbb{R}^d$

	Strongly convex?	**Smooth?**	**Lipschitz?**
Exponential loss (in $\mathbb{R}$)	NO	NO	NO
Hinge loss (in $\mathbb{R}$)	NO	NO	YES
Logistic loss (in $\mathbb{R}$)	NO	YES	YES

However, we typically only need the domain to be a compact set of $\mathbb{R}$