From Convex Analysis to Learning, Prediction, and Elicitation[*]
Lecture 10: Proper Scoring Rules and Revelation Principle

Lunjia Hu

1 Proper Scoring Rules

Definition 1. *We say a scoring rule $s : [0,1] \times \{0,1\} \to \mathbb{R}$ is* proper *if for every $p, q \in [0,1]$,*

$$\mathbb{E}_{y \sim p} s(p,y) \geq \mathbb{E}_{y \sim p} s(q,y).$$

Theorem 1. *Let $s : [0,1] \times \{0,1\} \to \mathbb{R}$ be a proper scoring rule. There exists a convex function $\varphi : [0,1] \to \mathbb{R}$ and its subgradient $\nabla\varphi : [0,1] \to \mathbb{R}$ such that*

$$s(q,y) = \varphi(q) + (y-q)\nabla\varphi(q) \quad \text{for every } q \in [0,1] \text{ and } y \in \{0,1\}. \tag{1}$$

Proof. We extend the domain of s from $[0,1] \times \{0,1\}$ to $[0,1] \times [0,1]$ as follows: for every $p, q \in [0,1]$, define $s(q,p) := \mathbb{E}_{y \sim p} s(q,y)$. Define $\varphi(p) := s(p,p)$. By the definition of properness,

$$\varphi(p) \geq s(q,p) \quad \text{for every } p, q \in [0,1]. \tag{2}$$

Let us consider a fixed $q \in [0,1]$. The function $s(q,p)$ is affine in p, and when $p = q$, the inequality (2) becomes an equality. Therefore, the graph of $s(q,p)$ (as an affine function of p) is "tangent" to the graph of $\varphi(p)$ at $p = q$. In particular, the slope of the graph of $s(q,p)$ is a subgradient of φ at q. Let $\nabla\varphi(q)$ denote that subgradient. Since the subgradient exists for every $q \in [0,1]$, we know that φ is convex. Moreover,

$$s(q,p) = \varphi(q) + (p-q)\nabla\varphi(q) \quad \text{for every } p, q \in [0,1].$$

This implies (1) as a special case. □

Remark 1. *Let $\psi : \mathbb{R} \to \mathbb{R}$ be the convex conjugate of φ. We have*

$$0 = \Gamma_{\varphi,\psi}(q, \nabla\varphi(q)) = \varphi(q) + \psi(\nabla\varphi(q)) - q\nabla\varphi(q).$$

Plugging this into (1), we have

$$s(q,y) = y\nabla\varphi(q) - \psi(\nabla\varphi(q)).$$

In particular, for every fixed y, the function $s(q,y)$ is concave in $\nabla\varphi(q)$, though it is not necessarily concave in q itself.

[*] `https://lunjiahu.com/convex-analysis/`

Remark 2. *Definition 1 and Theorem 1 extends beyond binary outcomes as follows.*

Definition 2. *Suppose there are k possible outcomes $y = 1, \ldots, k$. We say a scoring rule $s : \Delta_k \times [k] \to \mathbb{R}$ is* proper *if for every $p, q \in \Delta_k$,*

$$\mathbb{E}_{y \sim p} s(p, y) \geq \mathbb{E}_{y \sim p} s(q, y).$$

Theorem 2. *Let $s : \Delta_k \times [k] \to \mathbb{R}$ be a proper scoring rule. There exists a convex function $\varphi : \Delta_k \to \mathbb{R}$ and its subgradient $\nabla\varphi : \Delta_k \to \mathbb{R}^k$ such that*

$$s(q, y) = \varphi(q) + \langle \mathbf{e}_y - q, \nabla\varphi(q) \rangle \quad \text{for every } q \in \Delta_k \text{ and } y \in [k].$$

Here $\mathbf{e}_y$ is the unit vector with its y-th coordinate being 1 (and all other coordinates being 0).

Example 1 (Cross-entropy Loss)**.** *The cross-entropy loss $-s(q, y) = -\ln q_y$ is obtained by choosing the convex function φ as the negative Shannon entropy:*

$$\begin{aligned}
\varphi(q) &= \sum_{i=1}^{k} q_i \ln q_i, \\
\nabla\varphi(q) &= (\ln q_1, \ldots, \ln q_k), \\
s(q, y) &= \varphi(q) + \langle \mathbf{e}_y - q, \nabla\varphi(q) \rangle = \ln q_y, \\
-s(q, y) &= -\ln q_y.
\end{aligned}$$

Example 2 (Squared loss, a.k.a Brier loss))**.** *The squared loss $-s(q, y) = \frac{1}{2}\|\mathbf{e}_y - q\|_2^2$ is obtained by choosing φ as folllows:*

$$\begin{aligned}
\varphi(q) &:= \frac{1}{2}(\|q\|_2^2 - 1) \\
\nabla\varphi(q) &= q, \\
s(q, y) &= \varphi(q) + \langle \mathbf{e}_y - q, \nabla\varphi(q) \rangle = q_y - \frac{1}{2}\|q\|_2^2 - \frac{1}{2} = -\frac{1}{2}\|\mathbf{e}_y - q\|_2^2, \\
-s(q, y) &= \frac{1}{2}\|\mathbf{e}_y - q\|_2^2.
\end{aligned}$$

2 Revelation Principle

Theorem 3. *Let $u : A \times [k] \to \mathbb{R}$ be an arbitrary function. Define best response function $r_u : \Delta_k \to A$ such that $r_u(q) = \arg\max_{a \in A} \mathbb{E}_{y \sim q} u(a, y)$. Define*

$$s(q, y) := u(r_u(q), y).$$

Then s is proper.

Example 3 (Classification error)**.** *Suppose $A = [k]$, and $u(a, y) = \mathbb{I}[a = y]$. We have $r_u(q) = \arg\max_{a \in [k]} q_a$. We obtain the following proper scoring rule:*

$$s(q, y) = u(r_u(q), y) = \mathbb{I}[y = \arg\max_{a \in [k]} q_a].$$

3 V-shape Decomposition

Theorem 4. *Let $\varphi : [0,1] \to \mathbb{R}$ be a twice-differentiable convex function. Then*

$$\varphi'(v) = \varphi'(0) + \int_0^1 \varphi''(t)\mathbb{I}[v - t \geq 0]\mathrm{d}t,$$

$$\varphi(v) = \varphi(0) + \varphi'(0)v + \int_0^1 \varphi''(t)\max\{v - t, 0\}\mathrm{d}t.$$

4 Generalized Linear Models

Learning a generalized linear model. Let D be a distribution of $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ that we wish to learn. Assume $(x, y) \sim D$ satisfies $y = \sigma(\langle a^*, x\rangle) + z$, where $a^* \in \mathbb{R}^d$ is the ground-truth parameter, $\sigma : \mathbb{R} \to \mathbb{R}$ is a monotonically increasing *link* function, and $z \in \mathbb{R}$ is random mean-zero noise independent of x. Our goal is to learn a^* assuming knowledge of σ.

Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a loss function. We can estimate a^* by finding $a \in \mathbb{R}^d$ that minimizes

$$L(a) := \sum_{i=1}^n \ell(\sigma(\langle a, x_i\rangle), y_i)$$

over i.i.d. examples $(x_1, y_1), \ldots, (x_n, y_n)$ drawn from D.

The question is how we should choose the loss function ℓ. Ideally, the choice of ℓ should make L convex in a, and $\mathbb{E}[L(a)]$ should be minimized when $a = a^*$.

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\sigma(t) = \nabla\psi(t)$. Let φ be the convex conjugate of ψ. We define

$$\ell(q, y) := -\varphi(q) - (y - q)\nabla\varphi(q).$$

This is a proper loss function: when y is drawn from a distribution with mean p, the expected loss $\mathbb{E}[\ell(q, y)]$ is minimized at $q = p$. Consequently, when $y = \sigma(\langle a^*, x\rangle) + z$ for a mean-zero noize z, the expected loss $\mathbb{E}[\ell(\sigma(\langle a, x\rangle), y)]$ is minimized at $a = a^*$. Note that for $q, t \in \mathbb{R}$ such that $q = \sigma(t) = \nabla\psi(t)$, we have

$$\ell(q, y) = \psi(t) - yt.$$

Therefore,

$$\ell(\sigma(\langle a, x\rangle), y) = \psi(\langle a, x\rangle) - y\langle a, x\rangle$$

is a convex function of a.