# Proper scoring rules

Johanna Ziegel

ETH Zurich

CUSO Summer School
7–10 September 2025

# Introduction: Probabilistic predictions

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Amount of precipitation tomorrow in Cambridge and Oxford. ($Y \in \mathcal{Y} = \mathbb{R}^2$)
- Single valued "best guess" $z \in \mathcal{Y}$ does not quantify uncertainty.

# Introduction: Probabilistic predictions

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Amount of precipitation tomorrow in Cambridge and Oxford. ($Y \in \mathcal{Y} = \mathbb{R}^2$)
- Single valued "best guess" $z \in \mathcal{Y}$ does not quantify uncertainty.
- Better: Quantify uncertainty of $Y$ by a *probabilistic prediction* $F$.
  - $F$ is a distribution on $\mathcal{Y}$.

# Introduction: Probabilistic predictions

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Amount of precipitation tomorrow in Cambridge and Oxford. ($Y \in \mathcal{Y} = \mathbb{R}^2$)
- Single valued "best guess" $z \in \mathcal{Y}$ does not quantify uncertainty.
- Better: Quantify uncertainty of $Y$ by a *probabilistic prediction* $F$.
  - $F$ is a distribution on $\mathcal{Y}$.
- If $X$ is information available for prediction, $F$ should approximate $\mathcal{L}(Y \mid X)$.

# Introduction: Probabilistic predictions

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Amount of precipitation tomorrow in Cambridge and Oxford. ($Y \in \mathcal{Y} = \mathbb{R}^2$)
- Single valued "best guess" $z \in \mathcal{Y}$ does not quantify uncertainty.
- Better: Quantify uncertainty of $Y$ by a *probabilistic prediction* $F$.
  - $F$ is a distribution on $\mathcal{Y}$.
- If $X$ is information available for prediction, $F$ should approximate $\mathcal{L}(Y \mid X)$.
- Other possibilities to quantify uncertainty of $Y$: prediction intervals, predictions of some measure of variability, ...

# Introduction: Probabilistic predictions

- Let $Y \in \mathcal{Y}$ be an unknown future outcome.
  - Temperature tomorrow at 12:00 in Cambridge. ($Y \in \mathcal{Y} = \mathbb{R}$)
  - Event of rain tomorrow in London. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Default of credit card client. ($Y \in \mathcal{Y} = \{0, 1\}$)
  - Amount of precipitation tomorrow in Cambridge and Oxford. ($Y \in \mathcal{Y} = \mathbb{R}^2$)
- Single valued "best guess" $z \in \mathcal{Y}$ does not quantify uncertainty.
- Better: Quantify uncertainty of $Y$ by a *probabilistic prediction* $F$.
  - $F$ is a distribution on $\mathcal{Y}$.
- If $X$ is information available for prediction, $F$ should approximate $\mathcal{L}(Y \mid X)$.
- Other possibilities to quantify uncertainty of $Y$: prediction intervals, predictions of some measure of variability, ...
- Which loss functions can we use to compare probabilistic predictions?

Proper scoring rules.

# Outline

## Definition

- Motivation
- Definition
- Divergence and entropy
- Examples

# Motivation 1: Forecast comparison

Let $\mathcal{P}$ be a class of distributions on $\mathcal{Y}$.

Pick a loss function $S : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$, and compare average realized scores:
For $(F_1, G_1, Y_1), \ldots, (F_n, G_n, Y_n)$, let

$$\hat{S}_1 = \frac{1}{n} \sum_{i=1}^{n} S(F_i, Y_i) \quad \text{and} \quad \hat{S}_2 = \frac{1}{n} \sum_{i=1}^{n} S(G_i, Y_i).$$

The forecast with the smaller value $\hat{S}_j$ is better.

- When does this procedure make sense?
- $S$ needs to be a proper scoring rule.
- History: In meteorology, Brier (1950) showed that quadratic score for binary outcomes cannot be gamed.

# Motivation 2: McCarthy's forecasting tournament

- Forecasting agent makes probabilistic prediction $F$ for a random variable $Y$ and receives the penalty $S(F, Y)$.
- Rationally, if the agent believes $Y \sim G$ then it issues

$$\operatorname*{argmin}_F \mathbb{E}_G[S(F, Y)]$$

  $\mathbb{E}_G S(F, Y)$ means: take expectation with $Y \sim G$.
- Not necessarily equal to $G$.
- **McCarthy's idea.** Choose $S$ such that

$$G = \operatorname*{argmin}_F \mathbb{E}_G[S(F, Y)]$$

- McCarthy characterized such (differentiable) $S$ for $Y \in \{0, 1\}$ (McCarthy, 1956).
- Bregman divergences...$\sim$ 10 years before Bregman introduced them in 1967.

Let $\mathcal{P}$ be a convex class of distributions on $\mathcal{Y}$.

## Definition

A *scoring rule* is a function $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ that is suitably integrable.
A scoring rule $S$ is *proper* if

$$\mathbb{E}_F S(F, Y) \leq \mathbb{E}_F S(G, Y), \quad F, G \in \mathcal{P}, Y \sim F. \tag{1}$$

$S$ is *strictly* proper if equality implies $F = G$.

Let $\mathcal{P}$ be a convex class of distributions on $\mathcal{Y}$.

## Definition

A *scoring rule* is a function $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ that is suitably integrable.
A scoring rule $S$ is *proper* if

$$\mathbb{E}_F S(F, Y) \leq \mathbb{E}_F S(G, Y), \quad F, G \in \mathcal{P}, Y \sim F. \tag{1}$$

$S$ is *strictly* proper if equality implies $F = G$.

Equivalent to (1) is

$$F \in \operatorname*{argmin}_G \mathbb{E}_F S(G, Y) = \operatorname*{argmin}_G S(G, F).$$

Let $\mathcal{P}$ be a convex class of distributions on $\mathcal{Y}$.

## Definition

A *scoring rule* is a function $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ that is suitably integrable.
A scoring rule $S$ is *proper* if

$$\mathbb{E}_F S(F, Y) \leq \mathbb{E}_F S(G, Y), \quad F, G \in \mathcal{P}, Y \sim F. \tag{1}$$

$S$ is *strictly* proper if equality implies $F = G$.

Equivalent to (1) is

$$F \in \operatorname*{argmin}_G \mathbb{E}_F S(G, Y) = \operatorname*{argmin}_G S(G, F).$$

- Scoring rules are interpreted as penalties.
- Forecasts should be compared with proper scoring rules (Gneiting and Raftery, 2007).
- Proper scoring rules are also increasingly important in estimation (Dawid et al., 2016).
- New review article (Waghmare and Ziegel, 2025).

# Examples for $\mathcal{Y} = \{0, 1\}$

Distributions $F$ on $\mathcal{Y}$ can be identified with a parameter $p \in [0, 1]$.

## Brier score

$$S(p, y) = (y - p)^2, \quad p \in [0, 1],\ y \in \{0, 1\},$$

## Logarithmic score

$$S(p, y) = -y \log(p) - (1 - y) \log(1 - p), \quad p \in [0, 1],\ y \in \{0, 1\}.$$

# Divergence and entropy

For a scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$, we associate an *entropy*

$$H : \mathcal{P} \to \overline{R}, \quad H(F) = \int S(F, y) \, \mathrm{d}F(y) = \mathbb{E}_F S(F, Y) = S(F, F),$$

and a *divergence*

$$d : \mathcal{P} \times \mathcal{P} \to \overline{\mathbb{R}}, \quad d(F, G) = S(F, G) - H(G).$$

# Examples for $\mathcal{Y} = \mathbb{R}$

## Logarithmic Score (LogS)

$f$ density of $F$

$$\text{LogS}(F, y) = -\log f(y)$$

# Examples for $\mathcal{Y} = \mathbb{R}$

## Logarithmic Score (LogS)

$f$ density of $F$

$$\text{LogS}(F, y) = -\log f(y)$$

Entropy is Shannon entropy:

$$H(F) = -\int f(x) \log f(x) \, \mathrm{d}\mu(x) = -\mathbb{E}_F \log f(X)$$

# Examples for $\mathcal{Y} = \mathbb{R}$

## Logarithmic Score (LogS)

$f$ density of $F$

$$\text{LogS}(F, y) = -\log f(y)$$

Entropy is Shannon entropy:

$$H(F) = -\int f(x) \log f(x) \, \mathrm{d}\mu(x) = -\mathbb{E}_F \log f(X)$$

Divergence is Kullback-Leibler divergence: $g$ density of $G$

$$d(F, G) = \int g(y) \log \left( \frac{g(y)}{f(y)} \right) \mathrm{d}\mu(y) = D_{\mathrm{KL}}(G||F)$$

# Examples for $\mathcal{Y} = \mathbb{R}$

## Logarithmic Score (LogS)

$f$ density of $F$

$$\mathrm{LogS}(F, y) = -\log f(y)$$

Entropy is Shannon entropy:

$$H(F) = -\int f(x) \log f(x) \, \mathrm{d}\mu(x) = -\mathbb{E}_F \log f(X)$$

Divergence is Kullback-Leibler divergence: $g$ density of $G$

$$d(F, G) = \int g(y) \log \left( \frac{g(y)}{f(y)} \right) \mathrm{d}\mu(y) = D_{\mathrm{KL}}(G||F)$$

- Empirical risk minimization with respect to logarithmic score: Maximum likelihood estimation

# Examples for $\mathcal{Y} = \mathbb{R}$

## Continuous Ranked Probability Score (CRPS)

$F$ CDF, finite mean

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leq z\})^2 \, \mathrm{d}z$$

Entropy

$$H(F) = \int F(x)(1 - F(x)) \, \mathrm{d}x$$

Divergence $G$ CDF, finite mean

$$d(F, G) = \int (F(y) - G(y))^2 \, \mathrm{d}y$$

- Central role in forecast evaluation in meteorology.

# Estimation

# Motivation 3

Let $\mathcal{Y}$ be a Polish space and $\mathcal{X}$ be some measurable covariate space.

**Theorem**

*A scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is strictly proper if and only if for every pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ such that*

- *the conditional distributions $\mathbb{P}_{Y|X=x} \in \mathcal{P}$*

*we have*

$$\{\mathbb{P}_{Y|X=x}\}_x = \operatorname*{argmin}_{\{\mathbb{P}^x\}_x} \mathbb{E}\left[S(\mathbb{P}^X, Y)\right].$$

*where $\mathbb{P}^x : \mathcal{X} \to \mathcal{P}$.*

## Remark

- $$\mathbb{P}_{Y|X=\,\cdot\,} = \operatorname*{argmin}_{\mathbb{P}(\cdot)} \mathbb{E}\left[S(\mathbb{P}^X, Y)\right].$$

  is analogous to

$$\mathbb{E}[Y|X = \,\cdot\,] = \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[(Y - f(X))^2]$$

- Proper scoring rules are to conditional distributions what the squared error loss (and Bregman divergences) are to conditional expectations.

# Bias-Variance decomposition

## Theorem

- *Let* $\{\mathbb{P}_{\theta,x}\}_{\theta\in\Theta}$ *be a parametric family.*
- *Consider an estimator* $\mathbb{P}_{\hat{\theta},\cdot}$ *of* $\mathbb{P}_{Y|X=\cdot}$ *where* $\hat{\theta} = \hat{\theta}(\{(X_j, Y_j)\}_{j=1}^n)$.
- *Then,*

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \underbrace{\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X)]}_{\text{variance}} + \underbrace{\mathbb{E}[d(\bar{\mathbb{P}}_X, \mathbb{P}_{Y|X})]}_{\text{bias}}$$

*where*

$$\bar{\mathbb{P}}_x = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[d(\mathbb{P}_{\hat{\theta},x}, \mathbb{P})]$$

*and*

$$\mathbb{P}_{Y|X=x} = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[S(\mathbb{P}, Y) \mid X = x]$$

*is the best predictor, assuming the two exist.*

(Pfau, 2013)

# Analogy to classical bias-variance decomposition

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \underbrace{\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X)]}_{\text{variance}} + \underbrace{\mathbb{E}[d(\bar{\mathbb{P}}_X, \mathbb{P}_{Y|X})]}_{\text{bias}}$$

$$\bar{\mathbb{P}}_x = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[d(\mathbb{P}_{\hat{\theta},x}, \mathbb{P})] \qquad \mathbb{P}_{Y|X=x} = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[S(\mathbb{P}, Y) \mid X = x]$$

# Analogy to classical bias-variance decomposition

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \underbrace{\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X)]}_{\text{variance}} + \underbrace{\mathbb{E}[d(\bar{\mathbb{P}}_X, \mathbb{P}_{Y|X})]}_{\text{bias}}$$

$$\bar{\mathbb{P}}_x = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[d(\mathbb{P}_{\hat{\theta},x}, \mathbb{P})] \qquad \mathbb{P}_{Y|X=x} = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[S(\mathbb{P}, Y) \mid X = x]$$

Take $S(F, y) = (m(F) - y)^2$. Then

$$d(F, G) = \mathbb{E}_G(m(F) - Y)^2 - \operatorname{var}_G(Y) = (m(F) - m(G))^2.$$

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \mathbb{E}[\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X}) \mid X]] = \mathbb{E}(m_{\hat{\theta}}(X) - Y)^2 - \mathbb{E}\operatorname{var}(Y \mid X)$$

# Analogy to classical bias-variance decomposition

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \underbrace{\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X)]}_{\text{variance}} + \underbrace{\mathbb{E}[d(\bar{\mathbb{P}}_X, \mathbb{P}_{Y|X})]}_{\text{bias}}$$

$$\bar{\mathbb{P}}_x = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[d(\mathbb{P}_{\hat{\theta},x}, \mathbb{P})] \qquad \mathbb{P}_{Y|X=x} = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[S(\mathbb{P}, Y) \mid X = x]$$

Take $S(F, y) = (m(F) - y)^2$. Then

$$d(F, G) = \mathbb{E}_G(m(F) - Y)^2 - \operatorname{var}_G(Y) = (m(F) - m(G))^2.$$

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \mathbb{E}[\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X}) \mid X]] = \mathbb{E}(m_{\hat{\theta}}(X) - Y)^2 - \mathbb{E}\operatorname{var}(Y \mid X)$$

$$\operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[d(\mathbb{P}_{\hat{\theta},x}, \mathbb{P})] \equiv \operatorname*{argmin}_{z} \mathbb{E}(m_{\hat{\theta}}(x) - z)^2 = \mathbb{E}_{\hat{\theta}} m_{\hat{\theta}}(x).$$

# Analogy to classical bias-variance decomposition

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \underbrace{\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X)]}_{\text{variance}} + \underbrace{\mathbb{E}[d(\bar{\mathbb{P}}_X, \mathbb{P}_{Y|X})]}_{\text{bias}}$$

$$\bar{\mathbb{P}}_x = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[d(\mathbb{P}_{\hat{\theta},x}, \mathbb{P})] \qquad \mathbb{P}_{Y|X=x} = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[S(\mathbb{P}, Y) \mid X = x]$$

Take $S(F, y) = (m(F) - y)^2$. Then

$$d(F, G) = \mathbb{E}_G(m(F) - Y)^2 - \operatorname{var}_G(Y) = (m(F) - m(G))^2.$$

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \mathbb{E}[\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X}) \mid X]] = \mathbb{E}(m_{\hat{\theta}}(X) - Y)^2 - \mathbb{E}\operatorname{var}(Y \mid X)$$

$$\operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[d(\mathbb{P}_{\hat{\theta},x}, \mathbb{P})] \equiv \operatorname*{argmin}_{z} \mathbb{E}(m_{\hat{\theta}}(x) - z)^2 = \mathbb{E}_{\hat{\theta}} m_{\hat{\theta}}(x).$$

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X)] = \mathbb{E}[\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X) \mid X]] = \mathbb{E}(m_{\hat{\theta}}(X) - \mathbb{E}_{\hat{\theta}} m_{\hat{\theta}}(X))^2 \quad \text{variance}$$

# Analogy to classical bias-variance decomposition

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \underbrace{\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X)]}_{\text{variance}} + \underbrace{\mathbb{E}[d(\bar{\mathbb{P}}_X, \mathbb{P}_{Y|X})]}_{\text{bias}}$$

$$\bar{\mathbb{P}}_x = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[d(\mathbb{P}_{\hat{\theta},x}, \mathbb{P})] \qquad \mathbb{P}_{Y|X=x} = \operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[S(\mathbb{P}, Y) \mid X = x]$$

Take $S(F, y) = (m(F) - y)^2$. Then

$$d(F, G) = \mathbb{E}_G(m(F) - Y)^2 - \operatorname{var}_G(Y) = (m(F) - m(G))^2.$$

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X})] = \mathbb{E}[\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \mathbb{P}_{Y|X}) \mid X]] = \mathbb{E}(m_{\hat{\theta}}(X) - Y)^2 - \mathbb{E}\operatorname{var}(Y \mid X)$$

$$\operatorname*{argmin}_{\mathbb{P}} \mathbb{E}[d(\mathbb{P}_{\hat{\theta},x}, \mathbb{P})] \equiv \operatorname*{argmin}_{z} \mathbb{E}(m_{\hat{\theta}}(x) - z)^2 = \mathbb{E}_{\hat{\theta}} m_{\hat{\theta}}(x).$$

$$\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X)] = \mathbb{E}[\mathbb{E}[d(\mathbb{P}_{\hat{\theta},X}, \bar{\mathbb{P}}_X) \mid X]] = \mathbb{E}(m_{\hat{\theta}}(X) - \mathbb{E}_{\hat{\theta}} m_{\hat{\theta}}(X))^2 \quad \text{variance}$$

$$\mathbb{E}[d(\bar{\mathbb{P}}_X, \mathbb{P}_{Y|X})] = \mathbb{E}[\mathbb{E}[d(\bar{\mathbb{P}}_X, \mathbb{P}_{Y|X}) \mid X]] = \mathbb{E}(\mathbb{E}_{\hat{\theta}} m_{\hat{\theta}}(X) - \mathbb{E}(Y \mid X))^2 \quad \text{bias}$$

## Classes of scoring rules

- Characterization
- Local scoring rules
- $f$-scores and $gf$-scores
- Kernel scores

# Characterization of proper scoring rules

Let $\mathcal{P}$ be a convex class of distributions on $\mathcal{Y}$.

## Preliminaries

- Concavity: $H : \mathcal{P} \to \overline{\mathbb{R}}$ such that for $F, G \in \mathcal{P}$,

$$H(\alpha F + (1-\alpha)G) \geq \alpha H(F) + (1-\alpha)H(G)$$

  Entropies $H$ are concave!

- Supergradient: $h_F \colon \mathcal{Y} \to \overline{\mathbb{R}}$ that is $G$-integrable and

$$H(F) + \int h_F(y)\, \mathrm{d}(G - F)(y) \geq H(G)$$

  for all $G \in \mathcal{P}$.

- Regularity: A scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is called *regular* if $H(F) = S(F, F)$ is finite and $S(F, G) > -\infty$ for every $F, G \in \mathcal{P}$ with $F \neq G$.

## Theorem

*A regular scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is (strictly) proper if and only if there is a (strictly) concave function $H : \mathcal{P} \to \mathbb{R}$ such that*

$$S(F, y) = H(F) + h_F(y) - \int h_F(x)\, dF(x)$$

*for every $F \in \mathcal{P}$ and $y \in \mathcal{Y}$, where $h_F$ is a supergradient of $H$ at $F$.*

- If $H$ is differentiable, $h_F = \nabla_F H$, where $\nabla_F H$ is such that, for all $G \in \mathcal{P}$

$$\lim_{\alpha \downarrow 0} \frac{1}{\alpha} [H((1-\alpha)F + \alpha G) - H(F)] = \int \nabla_F H(y)\, \mathrm{d}G(y),$$

- and

$$d(F, G) = H(F) - H(G) - \int \nabla_F H(y)\, \mathrm{d}(F - G)(y).$$

(Gneiting and Raftery, 2007)

# Local scoring rules

- $\mathcal{Y} \subseteq \mathbb{R}^d$ open
- $\mathcal{P}^k$: probability measures on $\mathcal{Y}$ with $k$-times differentiable densities

## Definition

A proper scoring rule $S : \mathcal{P}^k \times \mathcal{Y} \to \overline{\mathbb{R}}$ is *local of order $k$* if

$$S(F, y) = s(y, f(y), \ldots, \nabla_y^k f(y)), \quad F \in \mathcal{P},$$

where $f$ is the density of $F$.

## Example

Local scoring rule of order 0: $S(F, y) = -\log f(y)$.

# Uniqueness of logarithmic score

## Theorem

*Logarithmic score is (essentially) only differentiable local scoring rule of order* 0.

# Uniqueness of logarithmic score

## Theorem

*Logarithmic score is (essentially) only differentiable local scoring rule of order* 0.

## Proof.

Let $S(F, y) = s(y, f(y))$ with $s(y, z)$ differentiable. Then $H(F) = \int s(y, f(y))f(y)\,\mathrm{d}y$.

$$\lim_{\alpha \downarrow 0} \frac{1}{\alpha}\left[H((1-\alpha)F + \alpha G) - H(F)\right] = \int \nabla_F H(y)\,\mathrm{d}G(y)$$

$$= \int \Big(s(y, f(y)) + \frac{\partial}{\partial z} s(x, f(y))f(y)\Big) g(y)\,\mathrm{d}y - \int s(y, f(y))f(y)\,\mathrm{d}y - \int \frac{\partial}{\partial z} s(y, f(y))(f(y))^2\,\mathrm{d}y$$

# Uniqueness of logarithmic score

## Theorem

*Logarithmic score is (essentially) only differentiable local scoring rule of order* 0.

## Proof.

Let $S(F,y) = s(y, f(y))$ with $s(y,z)$ differentiable. Then $H(F) = \int s(y, f(y)) f(y)\,\mathrm{d}y$.

$$\lim_{\alpha \downarrow 0} \frac{1}{\alpha}\left[H((1-\alpha)F + \alpha G) - H(F)\right] = \int \nabla_F H(y)\,\mathrm{d}G(y)$$

$$= \int \Big(s(y, f(y)) + \frac{\partial}{\partial z} s(x, f(y)) f(y)\Big) g(y)\,\mathrm{d}y - \int s(y, f(y)) f(y)\,\mathrm{d}y - \int \frac{\partial}{\partial z} s(y, f(y)) (f(y))^2\,\mathrm{d}y$$

Hence, by the characterization theorem

$$S(F,y) = s(y, f(y)) + f(y)\frac{\partial}{\partial y} s(y, f(y)) - \int \underbrace{\frac{\partial}{\partial z} s(y, f(w))(f(w))^2}_{= c f(w)}\,\mathrm{d}w$$

$$\Rightarrow \frac{\partial}{\partial z} s(y, f(w)) = \frac{c}{f(w)} \Rightarrow s(y, f(y)) = a \log f(y) + b \text{ for some } a, b \in \mathbb{R}. \qquad \square$$

# Score matching

## Example (Hyvärinen Score)

Let $\mathcal{Y} = \mathbb{R}^d$ and $\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})\| \to 0$ as $\|\boldsymbol{x}\| \to \infty$ for $p \in \mathcal{P}^2$. Then

$$S(P, \boldsymbol{y}) = \Delta_{\boldsymbol{y}} \log p(\boldsymbol{y}) + \frac{1}{2} \|\nabla_{\boldsymbol{y}} \log p(\boldsymbol{y})\|^2 = \sum_j \left\{ \frac{\partial^2 \log p}{\partial y_j^2} + \frac{1}{2} \left[ \frac{\partial \log p}{\partial y_j} \right]^2 \right\},$$

where $\nabla_{\boldsymbol{y}} f$ and $\Delta_{\boldsymbol{y}} f$ are gradient and Laplacian, is a strictly proper scoring rule.

# Score matching

## Example (Hyvärinen Score)

Let $\mathcal{Y} = \mathbb{R}^d$ and $\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})\| \to 0$ as $\|\boldsymbol{x}\| \to \infty$ for $p \in \mathcal{P}^2$. Then

$$S(P, \boldsymbol{y}) = \Delta_{\boldsymbol{y}} \log p(\boldsymbol{y}) + \frac{1}{2}\|\nabla_{\boldsymbol{y}} \log p(\boldsymbol{y})\|^2 = \sum_j \left\{ \frac{\partial^2 \log p}{\partial y_j^2} + \frac{1}{2}\left[\frac{\partial \log p}{\partial y_j}\right]^2 \right\},$$

where $\nabla_{\boldsymbol{y}} f$ and $\Delta_{\boldsymbol{y}} f$ are gradient and Laplacian, is a strictly proper scoring rule.

## Non-normalized Densities

Consider the parametric family $\{P_\theta : \theta \in \Theta\}$. Let $dP_\theta/d\mu = p_\theta(\boldsymbol{x}) \propto \exp \eta_\theta(\boldsymbol{x})$.

$$S(P_\theta, \boldsymbol{x}) = \Delta_{\boldsymbol{x}} \eta_\theta(\boldsymbol{x}) + \frac{1}{2}\|\nabla_{\boldsymbol{x}} \eta_\theta(\boldsymbol{x})\|^2$$

# Score matching

## Example (Hyvärinen Score)

Let $\mathcal{Y} = \mathbb{R}^d$ and $\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})\| \to 0$ as $\|\boldsymbol{x}\| \to \infty$ for $p \in \mathcal{P}^2$. Then

$$S(P, \boldsymbol{y}) = \Delta_{\boldsymbol{y}} \log p(\boldsymbol{y}) + \frac{1}{2}\|\nabla_{\boldsymbol{y}} \log p(\boldsymbol{y})\|^2 = \sum_j \left\{ \frac{\partial^2 \log p}{\partial y_j^2} + \frac{1}{2}\left[\frac{\partial \log p}{\partial y_j}\right]^2 \right\},$$

where $\nabla_{\boldsymbol{y}} f$ and $\Delta_{\boldsymbol{y}} f$ are gradient and Laplacian, is a strictly proper scoring rule.

## Non-normalized Densities

Consider the parametric family $\{P_\theta : \theta \in \Theta\}$. Let $dP_\theta/d\mu = p_\theta(\boldsymbol{x}) \propto \exp \eta_\theta(\boldsymbol{x})$.

$$S(P_\theta, \boldsymbol{x}) = \Delta_{\boldsymbol{x}} \eta_\theta(\boldsymbol{x}) + \frac{1}{2}\|\nabla_{\boldsymbol{x}} \eta_\theta(\boldsymbol{x})\|^2$$

We have

$$\min_\theta \mathbb{E}_Q S(P_\theta, \mathbf{Y}) = \min_\theta \mathbb{E}_Q \|\nabla_{\boldsymbol{y}} \log p_\theta(\mathbf{Y}) - \nabla_{\boldsymbol{y}} \log q(\mathbf{Y})\|^2$$

# $f$-scores

$\mu$ measure on $\mathcal{Y}$, $\mathcal{P}$ all absolutely continuous measures wrt $\mu$
$f : [0, \infty) \to \overline{\mathbb{R}}$ concave and differentiable. Define concave entropy

$$H_f(P) = \int f(p(x)) \, \mathrm{d}\mu(x), \quad P \in \mathcal{P};$$

$p$ density of $P \in \mathcal{P}$.

# $f$-scores

$\mu$ measure on $\mathcal{Y}$, $\mathcal{P}$ all absolutely continuous measures wrt $\mu$
$f : [0, \infty) \to \overline{\mathbb{R}}$ concave and differentiable. Define concave entropy

$$H_f(P) = \int f(p(x)) \, \mathrm{d}\mu(x), \quad P \in \mathcal{P};$$

$p$ density of $P \in \mathcal{P}$.
This yields the proper scoring rule

$$S_f(P, y) = f'(p(y)) + H_f(P) - \int f'(p(x))p(x) \, \mathrm{d}\mu(x).$$

# $f$-scores

$\mu$ measure on $\mathcal{Y}$, $\mathcal{P}$ all absolutely continuous measures wrt $\mu$
$f : [0, \infty) \to \overline{\mathbb{R}}$ concave and differentiable. Define concave entropy

$$H_f(P) = \int f(p(x)) \, \mathrm{d}\mu(x), \quad P \in \mathcal{P};$$

$p$ density of $P \in \mathcal{P}$.
This yields the proper scoring rule

$$S_f(P, y) = f'(p(y)) + H_f(P) - \int f'(p(x))p(x) \, \mathrm{d}\mu(x).$$

- Different name: Separable Bregman scores (Grünwald and Dawid, 2004)
- Only scoring rules of the form $S(P, y) = r(p(y)) + \ell(P)$.

# $f$-scores

$$S_f(P, y) = f'(p(y)) + H_f(P) - \int f'(p(x))p(x)\,\mathrm{d}\mu(x).$$

## Examples

| $f(u)$ | $S(P, y)$ | |
|---|---|---|
| $-u \log u$ | $-\log p(y)$ | logarithmic score |
| $-u^2$ | $-2p(y) + \int p(x)^2\,\mathrm{d}\mu(x)$ | quadratic score, Brier score |

# $gf$-scores

Consider entropies of the form

$$H(P) = g\left(\int f(p(x))\,\mathrm{d}\mu(x)\right) = g(H_f(P)),$$

$p$ density of $P \in \mathcal{P}$,

## $gf$-scores

Consider entropies of the form

$$H(P) = g\left(\int f(p(x))\,\mathrm{d}\mu(x)\right) = g(H_f(P)),$$

$p$ density of $P \in \mathcal{P}$,
where $f$, $g$ are such that $H$ is concave, for example

- $f$ concave, $g$ concave and increasing;
- $f$ convex, $g$ concave and decreasing.

If $f$, $g$ are differentiable, we obtain the proper scoring rule

$$S_{gf}(P, y) = g'(H_f(P))f'(p(y)) + g(H_f(P)) - g'(H_f(P)) \int f'(p(x))p(x)\,\mathrm{d}\mu(x).$$

## $gf$-scores

Consider entropies of the form

$$H(P) = g\left(\int f(p(x))\,\mathrm{d}\mu(x)\right) = g(H_f(P)),$$

$p$ density of $P \in \mathcal{P}$,
where $f$, $g$ are such that $H$ is concave, for example

- $f$ concave, $g$ concave and increasing;
- $f$ convex, $g$ concave and decreasing.

If $f$, $g$ are differentiable, we obtain the proper scoring rule

$$S_{gf}(P, y) = g'(H_f(P))f'(p(y)) + g(H_f(P)) - g'(H_f(P)) \int f'(p(x))p(x)\,\mathrm{d}\mu(x).$$

**Example (Spherical score)**

$f(u) = u^2$, $g(u) = -\sqrt{u}$, $S(P, y) = -p(y)/\sqrt{\int p(x)^2\,\mathrm{d}\mu(x)}$.

# Kernel scores: Motivation

$\mathcal{P}$: class of probability measures on $\mathbb{R}$ with finite mean.
Probability measures specified as CDFs $F$.

# Kernel scores: Motivation

$\mathcal{P}$: class of probability measures on $\mathbb{R}$ with finite mean.
Probability measures specified as CDFs $F$.

## Continuous Ranked Probability Score (CRPS)

$$\begin{aligned} S(F, y) &= \int_{\mathbb{R}} \left(F(x) - \mathbb{1}\{y \leq x\}\right)^2 \, \mathrm{d}(x) \\ &= \int_0^1 \left(\mathbb{1}\{y \leq F^{-1}(\alpha)\} - \alpha\right) \left(F^{-1}(\alpha) - y\right) \, \mathrm{d}\alpha \\ &= \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'| \end{aligned}$$

- Allows to compare discrete, continuous and mixed discrete-continuous distributions.
- Is becoming increasingly popoular also in estimation.

# Kernel scores: Motivation

Let $h(x, y) = |x - y|$.

## Transformation Models

Let $g_\theta(\mathbf{N}) \sim \mathbb{P}_\theta$.

$$\hat{S} = \frac{1}{m} \sum_{i=1}^{m} h(g_\theta(\mathbf{N}_i), y) - \frac{1}{2m(m-1)} \sum_{i=1}^{m} \sum_{i':i' \neq i} h(g_\theta(\mathbf{N}_i), g_\theta(\mathbf{N}_{i'}))$$

# Kernel scores: Motivation

Let $h(x, y) = |x - y|$.

## Transformation Models

Let $g_\theta(\mathbf{N}) \sim \mathbb{P}_\theta$.

$$\hat{S} = \frac{1}{m} \sum_{i=1}^{m} h(g_\theta(\mathbf{N}_i), y) - \frac{1}{2m(m-1)} \sum_{i=1}^{m} \sum_{i':i' \neq i} h(g_\theta(\mathbf{N}_i), g_\theta(\mathbf{N}_{i'}))$$

## Conditional Transformation Models

Let $g_\theta(\boldsymbol{x}, \mathbf{N}) \sim \mathbb{P}_{\theta,\boldsymbol{x}}$.

$$\hat{S} = \frac{1}{m} \sum_{i=1}^{m} h(g_\theta(\boldsymbol{x}, \mathbf{N}_i), y) - \frac{1}{2m(m-1)} \sum_{i=1}^{m} \sum_{i':i' \neq i} h(g_\theta(\boldsymbol{x}, \mathbf{N}_i), g_\theta(\boldsymbol{x}, \mathbf{N}_{i'}))$$

(Gneiting et al., 2005; Hothorn et al., 2014; Bouchacourt et al., 2016)

# Kernel scores

Instead of $h(x, y) = |x - y| \ldots$

# Kernel scores

Instead of $h(x, y) = |x - y| \ldots$

Take conditionally negative definite kernel $h$ on $\mathcal{Y}$:

# Kernel scores

Instead of $h(x, y) = |x - y| \ldots$

Take conditionally negative definite kernel $h$ on $\mathcal{Y}$: That is, $h : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is

- symmetric: $h(x, y) = h(y, x)$;
- $\forall n \geq 1$, $x_1, \ldots, x_n \in \mathcal{Y}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ with $\sum_{j=1}^{n} \alpha_j = 0$, we have

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j h(x_i, x_j) \leq 0.$$

Then,

$$H(P) = \frac{1}{2} \int \int h(x, y) \, \mathrm{d}P(x) \, \mathrm{d}P(y).$$

is concave with supergradient

$$\nabla_P H(y) = \int h(x, y) \, \mathrm{d}P(x) - 2H(P).$$

Let $\mathcal{P}_h = \{P \mid H(P) < \infty\}$.

**Theorem (Kernel score)**

*If $h : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is a (strongly) conditionally negative definite kernel, then $S_h : \mathcal{P}_h \times \mathcal{Y} \to \overline{\mathbb{R}}$ given by*

$$S_h(P, y) = \int h(x, y)\, dP(x) - \frac{1}{2} \int \int h(x, y)\, dP(x)\, dP(y)$$

*is a (strictly) proper scoring rule.*

Let $\mathcal{P}_h = \{P \mid H(P) < \infty\}$.

## Theorem (Kernel score)

*If $h : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is a (strongly) conditionally negative definite kernel, then $S_h : \mathcal{P}_h \times \mathcal{Y} \to \overline{\mathbb{R}}$ given by*

$$S_h(P, y) = \int h(x, y)\, dP(x) - \frac{1}{2} \int \int h(x, y)\, dP(x)\, dP(y)$$

*is a (strictly) proper scoring rule.*

- Divergence: $d(P, Q) = -\frac{1}{2} \int \int h(x, y)\, \mathrm{d}(P - Q)(x)\, \mathrm{d}(P - Q)(y)$
  Symmetric in $P$ and $Q$!
- $S(P, y) = d(P, \delta_y) + \frac{1}{2} h(y, y)$
  Divergence is itself a proper scoring rule.
- Divergence is squared maximum mean discrepancy (MMD).

(Gneiting and Raftery, 2007; Dawid, 2007; Steinwart and Ziegel, 2021)

# Connection to Hilbert space geometry

## Theorem

*There exists a Hilbert space $\mathcal{H}$ and a subset $\{\psi_x\}_{x\in\mathcal{Y}} \subset \mathcal{H}$ such that the divergence $d$ of $S$ satisfies*

$$d(P,Q) = -\frac{1}{2}\iint \|\psi_x - \psi_y\|^2_{\mathcal{H}}\, d(P-Q)(x)\, d(P-Q)(y).$$

*Moreover,*

$$d(P,Q) = \left\| \int \psi_x \, dP(x) - \int \psi_x \, dQ(x) \right\|^2_{\mathcal{H}}.$$

# Connection to Hilbert space geometry

## Theorem

*There exists a Hilbert space $\mathcal{H}$ and a subset $\{\psi_x\}_{x\in\mathcal{Y}} \subset \mathcal{H}$ such that the divergence $d$ of $S$ satisfies*

$$d(P,Q) = -\frac{1}{2}\iint \|\psi_x - \psi_y\|^2_{\mathcal{H}}\, d(P-Q)(x)\, d(P-Q)(y).$$

*Moreover,*

$$d(P,Q) = \left\| \int \psi_x \, dP(x) - \int \psi_x \, dQ(x) \right\|^2_{\mathcal{H}}.$$

Kernel score can be constructed on $\mathcal{Y}$ if

- we can construct a conditionally negative definite kernel on $\mathcal{Y}$;
- we can construct a positive definite kernel $k$ on $\mathcal{Y}$ (take $h = -k$);
- we can embed $\mathcal{Y}$ in a Hilbert space $\mathcal{H}$.

# Kernels on $\mathbb{R}^d$

## Bounded continuous kernels

Radial kernels that are strongly positive definite for any $d$:

$$k(x, y) = \varphi(\|x - y\|),$$

where

$$\varphi(t) = \int_0^\infty \exp(-t^2 s)\, \mathrm{d}\nu(s)$$

for a measure $\mu$ with $\operatorname{supp} \mu \neq \{0\}$.

(Sriperumbudur et al., 2011)

# Kernels on $\mathbb{R}^d$

## Bounded continuous kernels

Radial kernels that are strongly positive definite for any $d$:

$$k(x, y) = \varphi(\|x - y\|),$$

where

$$\varphi(t) = \int_0^\infty \exp(-t^2 s)\, \mathrm{d}\nu(s)$$

for a measure $\mu$ with $\operatorname{supp} \mu \neq \{0\}$.

(Sriperumbudur et al., 2011)

## Distance kernels

$$h(x, y) = \|x - y\|^\alpha,$$

for $\alpha \in (0, 2)$ are strongly conditionally negative definite for any $d$.

- CRPS corresponds to $d = 1$, $\alpha = 1$.
- Energy scores.

# Connection to energy distance

Székely and Rizzo (2004), Baringhaus and Franz (2004) introduced the *energy distance* between two distributions $P, Q$ on $\mathbb{R}^d$ with finite first moments

$$\mathbb{E}\|Z - W\| - \frac{1}{2}\mathbb{E}\|Z - Z'\| - \frac{1}{2}\mathbb{E}\|W - W'\|,$$

where $Z, Z', W, W'$ are independent with $Z, Z' \sim P$, $W, W' \sim Q$.

# Connection to energy distance

Székely and Rizzo (2004), Baringhaus and Franz (2004) introduced the *energy distance* between two distributions $P, Q$ on $\mathbb{R}^d$ with finite first moments

$$\mathbb{E}\|Z - W\| - \frac{1}{2}\mathbb{E}\|Z - Z'\| - \frac{1}{2}\mathbb{E}\|W - W'\|,$$

where $Z, Z', W, W'$ are independent with $Z, Z' \sim P$, $W, W' \sim Q$.

- Energy distance is the divergence of the kernel score with $h(x, y) = \|x - y\|$ called the *energy score*.
- Energy distance is a squared maximum mean discrepancy between $P$ and $Q$ (Sejdinovic et al., 2013).
- Energy score is a popular strictly proper scoring rule for multivariate outcomes.

# Distance kernels

When is

$$h(x, y) = \|x - y\|$$

strongly conditionally negative definite? Metric of strong negative type

# Distance kernels

When is

$$h(x, y) = \|x - y\|$$

strongly conditionally negative definite? Metric of strong negative type

- Separable Hilbert spaces
- Separable $L^p$-spaces for $1 < p \leq 2$

(Linde, 1986; Lyons, 2013; Sejdinovic et al., 2013)

# Characterizations of kernel scores

**Theorem**

*Let $S : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a proper scoring rule such that $\{\delta_y : y \in \mathcal{Y}\} \subseteq \mathcal{P}$ and the supergradient map $P \mapsto h_P$ is weakly continuous.*

*Then, the divergence $d$ of $S$ is symmetric if and only if $S$ is a kernel score.*

# Characterizations of kernel scores

**Theorem**

*Let $S : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a proper scoring rule such that $\{\delta_y : y \in \mathcal{Y}\} \subseteq \mathcal{P}$ and the supergradient map $P \mapsto h_P$ is weakly continuous.*

*Then, the divergence $d$ of $S$ is symmetric if and only if $S$ is a kernel score.*

- A proper scoring rule with $\{\delta_y : y \in \mathcal{Y}\} \subseteq \mathcal{P}$ corresponds to a squared metric on measures if and only if it is a kernel score!

(Waghmare and Ziegel, 2025)

# Characterizations of kernel scores

**Theorem**

*Let $S : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be a proper scoring rule such that $\{\delta_y : y \in \mathcal{Y}\} \subseteq \mathcal{P}$ and the supergradient map $P \mapsto h_P$ is weakly continuous.*

*Then, the divergence $d$ of $S$ is symmetric if and only if $S$ is a kernel score.*

- A proper scoring rule with $\{\delta_y : y \in \mathcal{Y}\} \subseteq \mathcal{P}$ corresponds to a squared metric on measures if and only if it is a kernel score!
- This also provides a natural motivation for using kernel-MMD for two sample testing.

(Waghmare and Ziegel, 2025)

# Characterizations of kernel scores

Let $S$ be a scoring rule.

- Translation invariance: For $\mathbf{y}, \mathbf{h} \in \mathbb{R}^d$,

$$S(P, \mathbf{y}) = S(P_{\mathbf{h}}, \mathbf{y} + \mathbf{h})$$

where $P_{\mathbf{h}}(A) = P(A + \mathbf{h})$ for Borel sets $A \subset \mathbb{R}^d$.

- Homogeneity: For every $c > 0$, $P \in \mathcal{P}$ and $\mathbf{y} \in \mathbb{R}^d$,

$$S(P_c, c\mathbf{y}) = c^{\alpha} S(P, \mathbf{y})$$

where $P_c(A) = P(c^{-1}A)$ for Borel sets $A \subset \mathbb{R}^d$.

- Isotropy: For every rotation matrix $\mathbf{U} \in \mathrm{SO}(d)$, $P \in \mathcal{P}$ and $\mathbf{y} \in \mathbb{R}^d$,

$$S(P_{\mathbf{U}}, \mathbf{U}\mathbf{y}) = S(P, \mathbf{y})$$

where $P_{\mathbf{U}}(A) = P(\mathbf{U}^{\top} A)$ for Borel sets $A \subset \mathbb{R}^d$.

Up to positive multiplicative constants:

1. CRPS $\rightarrow$ only 1-homogeneous translation invariant kernel score on $\mathbb{R}$, and
2. Energy Scores $\rightarrow$ only homogeneous isotropic translation invariant kernel scores on $\mathbb{R}^d$.

(Waghmare and Ziegel, 2025)

# Foundations

## Where do our loss functions come from?

- Logarithmic score.
  Only proper scoring rule of the form $S(P, y) = s(y, p(y))$.
- Kernel Maximum Mean Discrepancy.
  Only proper scoring rules which admit point measures and have symmetric divergences.
- Squared error loss.
  Only Bregman divergence which is symmetric and isotropic.

## Forecast comparison

- Information sets
- Decompositions

# Forecast comparison

Pick a strictly proper scoring rule $S$, compare the average realized score:
For $(F_1, G_1, Y_1), \ldots, (F_n, G_n, Y_n)$, let

$$\hat{S}_1 = \frac{1}{n}\sum_{i=1}^{n} S(F_i, Y_i) \quad \text{and} \quad \hat{S}_2 = \frac{1}{n}\sum_{i=1}^{n} S(G_i, Y_i).$$

The forecast with the smaller value $\hat{S}_j$ is better.

- Formal tests for differences between expected scores available.
  (Diebold and Mariano, 1995; Giacomini and White, 2006; Lai et al., 2011; Henzi and Ziegel, 2022; Choe and Ramdas, 2024)
- If $(F_i, Y_i)$ are iid, classical (asymptotic) t-test can be used.

## Simulation example

Let $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,2)$ be independent, and

$$\mathbb{P}(Y = 1 \mid X_1, X_2) = \Phi(X_1 + X_2).$$

Predictions:

$$p^{(0)} = 1/2, \quad p^{(1)} = \Phi\Big(\frac{X_1}{\sqrt{3}}\Big), \quad p^{(2)} = \Phi\Big(\frac{X_2}{\sqrt{2}}\Big), \quad p^{(3)} = \Phi(X_1 + X_2).$$

$n = 200$.

| Prediction | Brier Score | Logarithmic score |
|---|---|---|
| $p_0$ | 0.250 | 0.693 |
| $p_1$ | 0.213 | 0.613 |
| $p_2$ | 0.167 | 0.499 |
| $p_3$ | 0.116 | 0.355 |

Does forecast comparison depend on the choice of the proper scoring rule $S$? Generally, yes.

- Finite samples
- Non-nested information sets
- Uncalibrated predicitions/misspecified models

(Patton, 2020; Ziegel et al., 2020)

Does forecast comparison depend on the choice of the proper scoring rule $S$? Generally, yes.

- Finite samples
- Non-nested information sets
- Uncalibrated predicitions/misspecified models

(Patton, 2020; Ziegel et al., 2020)

Can we avoid the choice of a proper scoring rule $S$?

- For binary outcomes, sometimes yes (Murphy diagrams)
- Otherwise, no.

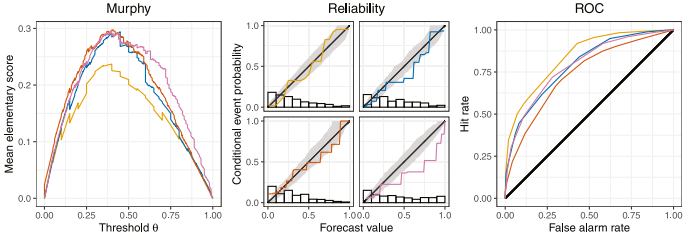
# Murphy diagrams

All proper scoring rules for binary outcomes $Y \in \{0, 1\}$ can be written as

$$S(p, y) = \int_0^1 S_\theta(p, y) \, \mathrm{d}H(\theta)$$

for some measure $H$ on $(0, 1)$.

- One-parameter family $(S_\theta(p, y))_\theta$ of elementary scores.
- Forecast $p$ is better with respect to all proper scoring rules $S$ if and only if it is better with respect to all $S_\theta$, $\theta \in (0, 1)$.




Ehm et al. (2016); Krüger and Ziegel (2021); Figure from Dimitriadis et al. (2024)

# Why should we use proper scoring rules to evaluate predictions?

- They incentivize truthful (calibrated) and informative forecasts.

Let $S$ be a (strictly) proper scoring rule.

**Theorem**

*Let $F = \mathcal{L}(Y \mid X)$, $G$ based on $X$ ($\sigma(X)$-measurable). Then,*

$$\mathbb{E}S(F, Y) \leq \mathbb{E}S(G, Y).$$

*Equality implies that $F = G$ almost surely.*

# Why should we use proper scoring rules to evaluate predictions?

- They incentivize truthful (calibrated) and informative forecasts.

Let $S$ be a (strictly) proper scoring rule.

**Theorem**

*Let $F = \mathcal{L}(Y \mid X)$, $G$ based on $X$ ($\sigma(X)$-measurable). Then,*

$$\mathbb{E}S(F, Y) \leq \mathbb{E}S(G, Y).$$

*Equality implies that $F = G$ almost surely.*

**Corollary**

*Let $F = \mathcal{L}(Y \mid X, Z)$, $G = \mathcal{L}(Y \mid X)$. Then,*

$$\mathbb{E}S(F, Y) \leq \mathbb{E}S(G, Y).$$

*Equality happens if and only if $Y$ and $Z$ are conditionally independent given $X$.*

(Holzmann and Eulert, 2014)

# Scoring rules and calibration

## Goal

Decompose

$$\hat{S} = \frac{1}{n} \sum_{k=1}^{n} S(F_i, y_i)$$

into interpretable terms quantifying miscalibration (MCB), discrimination ability (DSC), and uncertainty (UNC).

## Idea

$$\mathbb{E}S(F, Y) = \underbrace{\mathbb{E}S(F, Y) - \mathbb{E}S(\mathcal{L}(Y \mid F), Y)}_{\text{MCB}} - \underbrace{\left(\mathbb{E}S(\mathcal{L}(Y), Y) - \mathbb{E}S(\mathcal{L}(Y \mid F), Y)\right)}_{\text{DSC}} + \underbrace{\mathbb{E}S(\mathcal{L}(Y), Y)}_{\text{UNC}}$$

- $\mathcal{L}(Y \mid F)$ is best auto-calibrated forecast given information $F$.
- $\mathcal{L}(Y)$ is uninformative but auto-calibrated.

# Empirical translation of the idea

Data: $(F_1, y_1), \ldots, (F_n, y_n)$

# Empirical translation of the idea

Data: $(F_1, y_1), \ldots, (F_n, y_n)$

Uncertainty: $\mathrm{UNC} = \mathbb{E}S(\mathcal{L}(Y), Y)$

$$\overline{\mathrm{UNC}} = \frac{1}{n}\sum_{i=1}^{n} S\Big(\frac{1}{n}\sum_{k=1}^{n} \delta_{y_k}, y_i\Big)$$

# Empirical translation of the idea

Data: $(F_1, y_1), \ldots, (F_n, y_n)$

Uncertainty: $\text{UNC} = \mathbb{E}S(\mathcal{L}(Y), Y)$

$$\overline{\text{UNC}} = \frac{1}{n}\sum_{i=1}^{n} S\Big(\frac{1}{n}\sum_{k=1}^{n} \delta_{y_k}, y_i\Big)$$

Miscalibration: $\text{MCB} = \mathbb{E}S(F, Y) - \mathbb{E}S(\mathcal{L}(Y \mid F), Y)$

- Need estimate of $\mathcal{L}(Y \mid F)$ that is in-sample auto-calibrated and such that $\overline{\text{MCB}} \geq 0$. Generally not feasible.
- If $S = \text{CRPS}$, there is possible solution using isotonic distributional regression (Henzi et al., 2021).

Discrimination: $\text{DSC} = \mathbb{E}S(\mathcal{L}(Y), Y) - \mathbb{E}S(\mathcal{L}(Y \mid F), Y)$

- Estimate DSC by $\hat{S} - \overline{\text{MCB}} - \overline{\text{UNC}}$.

# Data application

Compare probabilistic quantitative precipitation forecasts.

## Numerical weather prediction models

- Physical model of the atmosphere is run with current (measured) inital conditions
- Initial conditions are measured with error: Several model runs with slightly perturbed inital conditions yields *ensemble of forecasts*
- Forecast ensembles are interpreted as random draws from the conditional distribution of the outcome
- Ensembles are usually biased and underdispersed: Statistical postprocessing

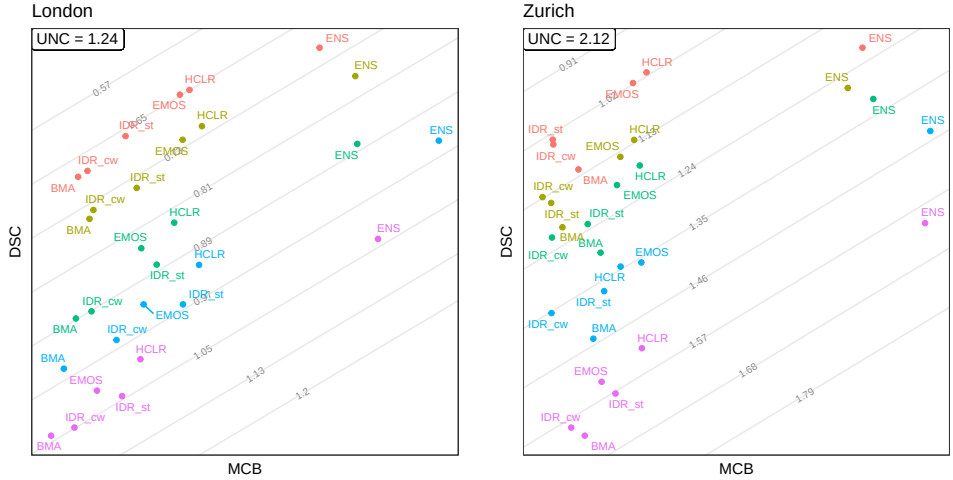
Bauer et al. (2015)

- Airport station observations at London and Zurich
- 52 member raw ensemble forecasts from European Centre for Medium-Range Weather Forecasts (ECMWF)
- Training data from 2007–2014, evaluation data from 2015-2016
- Prediction horizons of 1-5 days

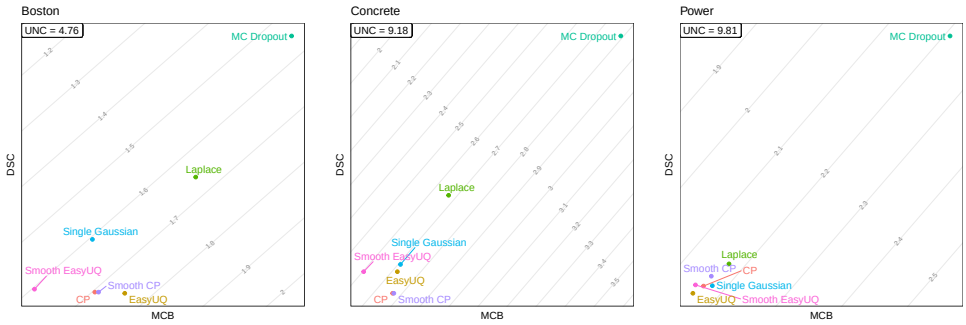
Postprocessing methods:

- Bayesian model averaging (BMA)
- Ensemble model output statistics (EMOS)
- Heteroscedastic censored logistic regression (HCLR)
- Isotonic distributional regression (IDR)

# Probabilistic quantitative precipitation forecasts



# Uncertainty quantification on benchmark datasets in machine learning




Arnold et al. (2024)

# Limitations

# Proper scoring rules and extremes

- Evaluation of forecast behaviour with respect to extreme events requires care (Lerch et al., 2017): Never condition on the outcomes when evaluating: Forecasters dilemma.

# Proper scoring rules and extremes

- Evaluation of forecast behaviour with respect to extreme events requires care (Lerch et al., 2017): Never condition on the outcomes when evaluating: Forecasters dilemma.
- Proper scoring rules are not necessarily directly useful: Expected scores cannot distinguish different tail behaviour (Brehmer and Strokorb, 2019).

# Proper scoring rules and extremes

- Evaluation of forecast behaviour with respect to extreme events requires care (Lerch et al., 2017): Never condition on the outcomes when evaluating: Forecasters dilemma.
- Proper scoring rules are not necessarily directly useful: Expected scores cannot distinguish different tail behaviour (Brehmer and Strokorb, 2019).
- However, evaluating calibration of probabilistic predictions with regards to tails is possible (Allen et al., 2025).

# Summary

# Summary

- Probabilistic predictions should be calibrated and sharp.
- Ideally, probabilistic predictions should be auto-calibrated.
- Proper scoring rules allow to compare probabilistic predictions simultaneously with respect to calibration and discrimination ability.

# References I


S. Allen, J. Koh, J. Segers, and J. Ziegel. Tail calibration of probabilistic forecasts. *Journal of the American Statistical Association*, 2025. To appear.

S. Arnold, E.-M. Walz, J. Ziegel, and T. Gneiting. Decompositions of the mean continuous ranked probability score. *Electronic Journal of Statistics*, 18:4992–5044, 2024.

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004.

P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 2015.

D. Bouchacourt, P. K. Mudigonda, and S. Nowozin. DISCO nets: Dissimilarity coefficient networks. In *Advances in Neural Information Processing Systems*, volume 29, pages 352–360, 2016. URL `https://papers.nips.cc/paper/6143-disco-nets-dissimilarity-coefficients-networks`.

J. Brehmer and K. Strokorb. Why scoring functions cannot assess tail properties. *Electronic Journal of Statistics*, 13:4015–4034, 2019.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

# References II


Y. J. Choe and A. Ramdas. Comparing sequential forecasters. *Operations Research*, 72: 1368–1387, 2024.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

A. P. Dawid, M. Musio, and L. Ventura. Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43:123–138, 2016.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13:253–263, 1995.

T. Dimitriadis, T. Gneiting, A. I. Jordan, and P. Vogel. Evaluating probabilistic classifiers: The triptych. *International Journal of Forecasting*, 40:1101–1122, 2024.

W. Ehm, T. Gneiting, A. Jordan, and F. Krüger. Of quantiles and expectiles: Consistent scoring functions, Choquet representations, and forecast rankings (with discussion). *Journal of the Royal Statistical Society: Series B*, 78:505–562, 2016.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74: 1545–1578, 2006.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

# References III


T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133:1098–1118, 2005.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

A. Henzi and J. F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, 109:647–663, 2022.

A. Henzi, J. F. Ziegel, and T. Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B*, 85:963–993, 2021.

H. Holzmann and M. Eulert. The role of the information set for forecasting – with applications to risk management. *Annals of Applied Statistics*, 8:595–621, 2014.

T. Hothorn, T. Kneib, and P. Bühlmann. Conditional transformation models. *Journal of the Royal Statistical Society: Series B*, 76:3–27, 2014.

F. Krüger and J. F. Ziegel. Generic conditions for forecast dominance. *Journal of Business & Economic Statistics*, 39:972–983, 2021.

# References IV


T. L. Lai, S. T. Gross, and D. B. Shen. Evaluating probability forecasts. *Annals of Statistics*, 39:2356–2382, 2011.

S. Lerch, T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32:106–127, 2017.

W. Linde. On Rudin's equimeasurability theorem for infinite dimensional Hilbert spaces. *Indiana University Mathematics Journal*, 35:235–243, 1986.

R. Lyons. Distance covariance in metric spaces. *Annals of Probability*, 41:3284–3305, 2013.

J. McCarthy. Measures of the value of information. *P. Natl. Acad. Sci. USA*, 42:654–655, 1956.

A. J. Patton. Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics*, 38:796–809, 2020.

D. Pfau. A generalized bias-variance decomposition for Bregman divergences. *Preprint, http://davidpfau.com/assets/generalized_bvd_proof.pdf*, 2013. Acessed 2025-02-25.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013.

# References V


B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

I. Steinwart and J. F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542, 2021.

G. J. Székely and M. Rizzo. Testing for equal distribution in high dimension. *InterStat*, 5, November 2004.

K. Waghmare and J. Ziegel. Proper scoring rules for estimation and forecast evaluation. *Annual Review of Statistics and Its Application*, 2025. To appear.

J. F. Ziegel, F. Krüger, A. Jordan, and F. Fasciati. Robust forecast evaluation of expected shortfall. *Journal of Financial Econometrics*, 18:95–120, 2020.