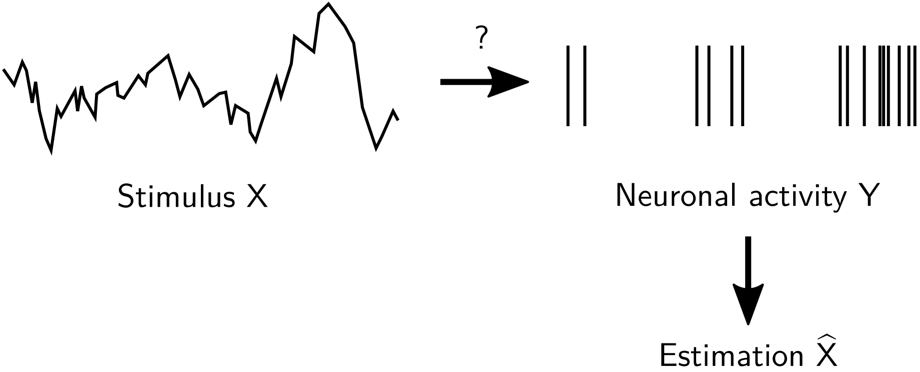
Optimal Activation Functions via Causal Wyner-Ziv Coding

Hui-An Shen
(based on joint work with
Jean-Pascal Pfister and Robert Graczyk)

France PhD Workshop on Information Theory

Efficient representation of stimuli and optimal estimation



E.g.
Sensory stimulus $\longrightarrow$ neuronal activity
Membrane potential $\longrightarrow$ action potential

Existing frameworks in neuroscience

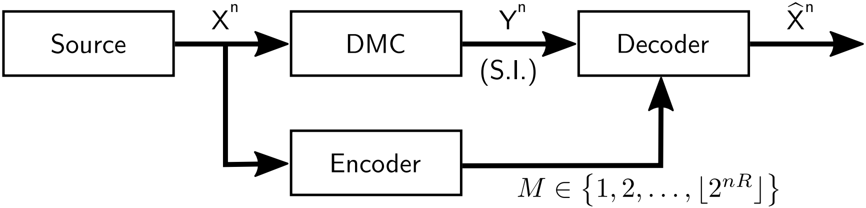
- Stimuli-generating distribution is completely known: optimal Bayes-estimation
- Restricted family of error measures (p-norms)
- Restricted family of activation functions (Gaussian, logistic function)

Objective

For a tolerable error (distortion value) under a specified distortion measure $\rho\colon \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_{\geq 0}$, minimize the "description rate" of the static r.v. X or stochastic process X_t.

The model I

- Source coding with causal side information at the decoder



- Continuous-time $\{X'_t\}$
  - $\{Y'_t\}$ inhomogeneous Poisson process of intensity $h'(X'_t)$
  - Let $X_i = X'_{i\Delta}$, $Y_i = Y'_{(i+1)\Delta} - Y'_{i\Delta}$
  - Y_i is Poisson-distributed with rate $h(X_i) \approx h'(X'_{i\Delta})\Delta$
    - Activation function $h\colon \mathcal{X} \to \mathbb{R}_{\geq 0}$

The model II

Encoder	$\mathcal{X}^n \to \{1, 2, \ldots, \lfloor 2^{nR} \rfloor\}$, $X^n \mapsto M$
Side information	causal pulse sequence Y^n
Decoder $\{dec_i\}$	$dec_1 \colon (M, Y_1) \mapsto \hat{X}_1$
	$dec_2 \colon (M, Y_1, Y_2) \mapsto \hat{X}_2$
	$\ldots$

- Reconstruction error:
  For $\rho \colon \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_{\geq 0}$, and threshold $D \geq 0$,
  require that $\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\rho(X_i, \hat{X}_i)\right] \leq D$.

$\longrightarrow$ Characterize the least description rate for general sources with causal SI: the *causal Wyner-Ziv problem*

The causal Wyner-Ziv rate-distortion function

$X_i, Y_i \overset{\text{iid}}{\sim} P_{XY}$ (source sequence $X^n \in \mathcal{X}^n$, side information sequence $Y^n \in \mathcal{Y}^n$)

Theorem (Weissmann and El Gamal 2006)

For $D \geq \mathbb{E}\left[\min_{\hat{x}} \rho(X, \hat{x})\right]$,

$$R(D) = \min_{\substack{P_{W|X}, f\colon (W,Y) \mapsto \hat{X} \\ \mathbb{E}\left[\rho(X,\hat{X})\right] \leq D}} I(X; W) \qquad (*)$$

where the expectation and MI are computed w.r.t.
$p_{XYW\hat{X}}(x, y, w, \hat{x}) = P_{XY}(x, y) P_{W|X}(w|x) \mathcal{I}\left\{\hat{x} = f(y, w)\right\}$.

Optimality criterion ("OPTA")

We define an optimal (neuronal) communication system as one operating at

1. arbitrarily close to $R(D)$
2. arbitrarily close to the capacity-cost function for the effective channel $P^*_{W|X}$ defined implicitly by the optimization.

$\longrightarrow$ Find h (optimal) that satisfies (1) and (2)

- Recall that Y_i is Poisson-distributed with rate $h(X_i)$,
  thus we identify $h\colon \mathcal{X} \to \mathbb{R}_{\geq 0}$ with the cost function,
  and cost $= \mathbb{E}\left[h(X)\right]$.

Results I: Optimal activation functions as relative entropy

Lemma (Gastpar, Rimoldi, and Vetterli 2003)

Let $c\colon \mathcal{X} \to \mathbb{R}_{\geq 0}$ be an input cost function. The input distribution P_X achieves the capacity of the channel $P_{W|X}$ at input cost $\mathbb{E}[c(X)]$ if and only if

$$c(x) = c_0 D_{\mathrm{KL}}\left(P_{W|X}(\cdot|x) \| P_W(\cdot)\right) + c_1 \quad (1)$$

whenever $P_X(x) > 0$, where $c_0 \geq 0$ and c_1 are constants.

And $P_W(\cdot) = \mathbb{E}[P_{W|X}(\cdot|X)]$ is the marginal distribution of W.

- At optimality, P_X is capacity-achieving for the channel $P^*_{W|X}$ implicitly defined by optimization (*)

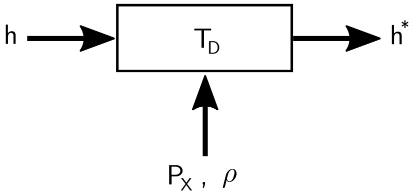
Results I (cont'd)

Theorem (S., Pfister, Graczyk 2024)

The optimal activation function h^* for the source distribution P_X, distortion function ρ, and distortion threshold D is given by

$$h^*(x) = c_0 D_{\mathrm{KL}}\left(P^*_{W|X}(\cdot|x) \| P^*_W(\cdot)\right) + c_1 \tag{2}$$

for suitable constants $c_0 \geq 0$ and c_1.



Results II: Existence of self-consistent optimal activation functions

Theorem (S., Pfister, Graczyk 2024)

Let $D_0 = \inf\{D \geq 0 : R(D) = 0\}$ be well-defined.
There exists a distortion value $D \in [\mathbb{E}[\min_{\hat{x}} \rho(X, \hat{x})], D_0)$ and a continuous activation function h s.t. $h^*(x) = h(x)$.

Proof: Apply Banach's fixed-point theorem to a complete metric space of (bounded) continuous functions $(C([0,1] \to \mathbb{R}), d_\infty)$.

Results III: The cost-distortion function

Definition (cost-distortion function)

Let $h^*_D \colon \mathcal{X} \to \mathbb{R}_{\geq 0}$ be the optimal activation function as defined in (2) with choice of constants $c_0 = 1$ and $c_1 = 0$. We define the cost-distortion function as

$$E^*(D) \triangleq \mathbb{E}\left[h^*_D(X)\right].$$

Theorem (S., Pfister, Graczyk 2024)

$$E^*(D) = R(D).$$

Proof:

$$\begin{aligned} E^*(D) &= \mathbb{E}\left[h^*_D(X)\right] \\ &= \mathbb{E}\left[D_{\mathrm{KL}}\left(P^*_{W|X}(\cdot|X) \| P^*_W(\cdot)\right)\right] \\ &= I(P_X; P^*_{W|X}) = R(D). \end{aligned}$$

Conclusion and future work

- Novel theoretical framework for the computation of optimal activation function
  - Action potentials as side information
  - Optimal activation function depends on tolerable distortion and takes the form of relative entropy

$$h^*(x) = c_0 D_{\mathrm{KL}}\left(P^*_{W|X}(\cdot|x) \| P^*_W(\cdot)\right) + c_1$$

- Generalize to functions beyond bounded continuous real functions
- Compute examples for $h^* \in C([0,1] \to \mathbb{R})$ and compare to biological activation functions