Post-hoc calibration

through lens of optimal transport and scoring rules

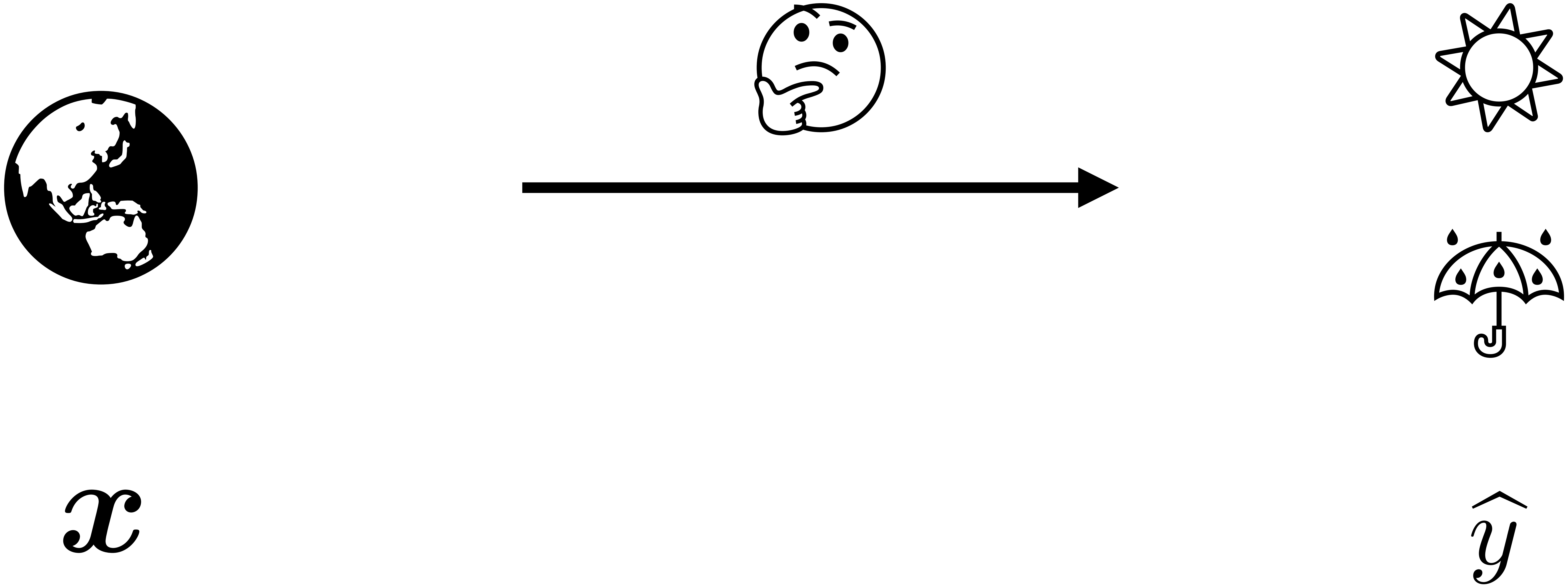
Han Bao

March 2nd, 2026

FIMI2026@Tokyo

Research Organization of Information and Systems
The Institute of Statistical Mathematics

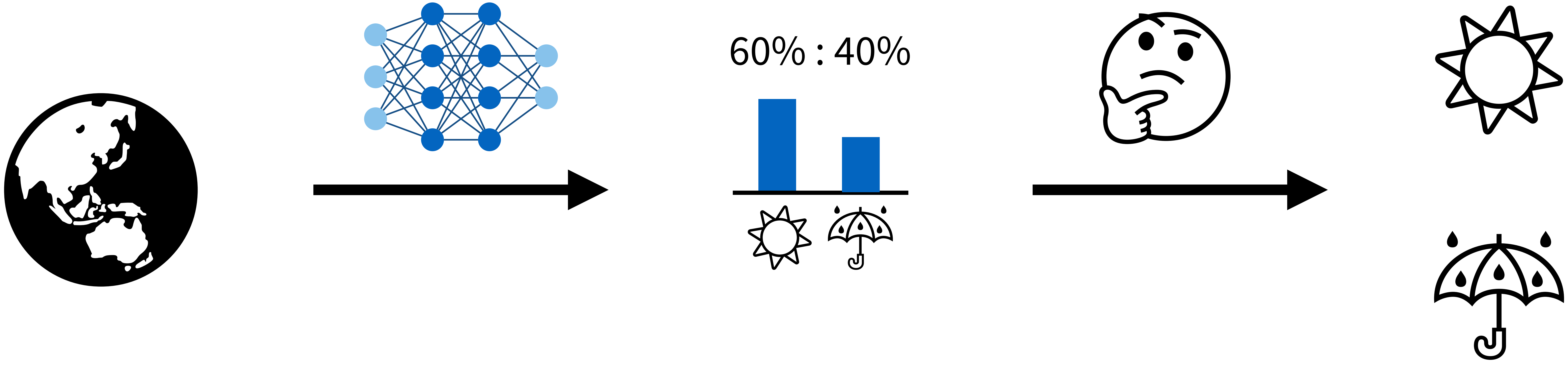
Classification



x

$\widehat{y}$

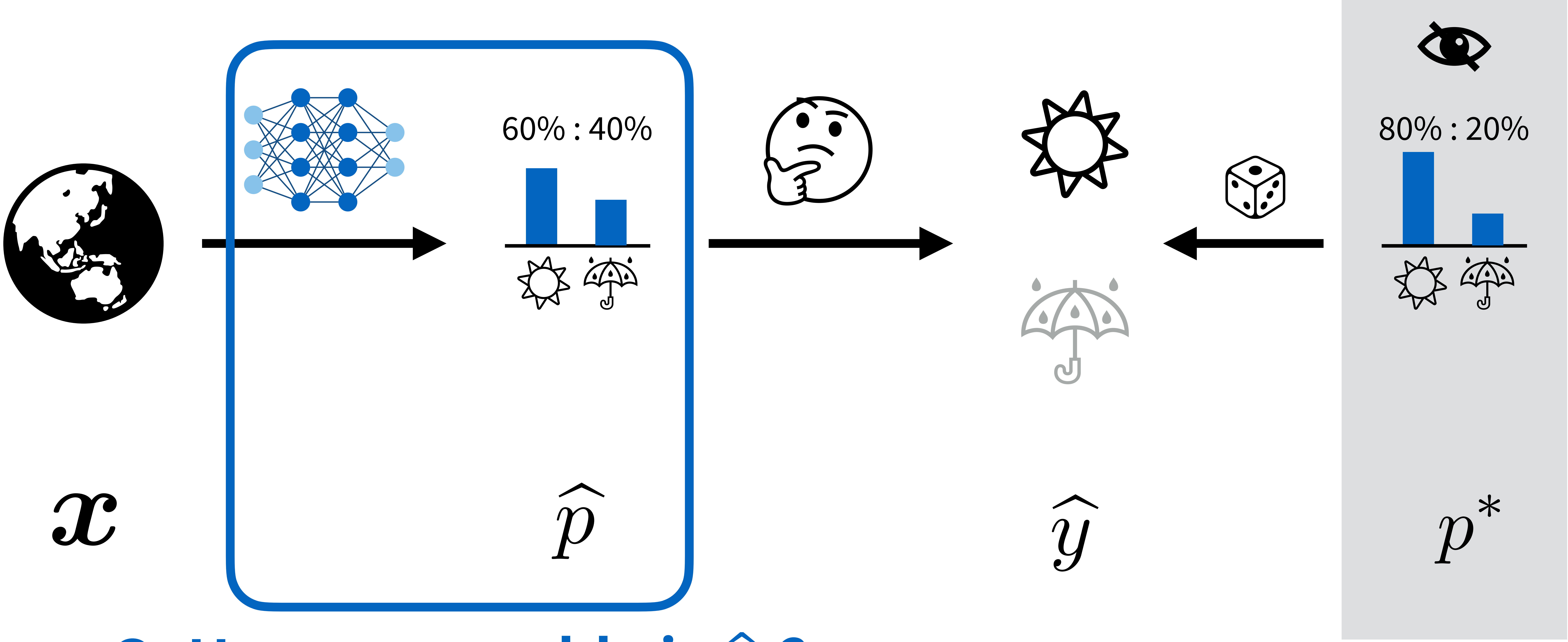
Classification with probabilistic prediction



60% : 40%

$\boldsymbol{x}$ $\widehat{p}$ $\widehat{y}$

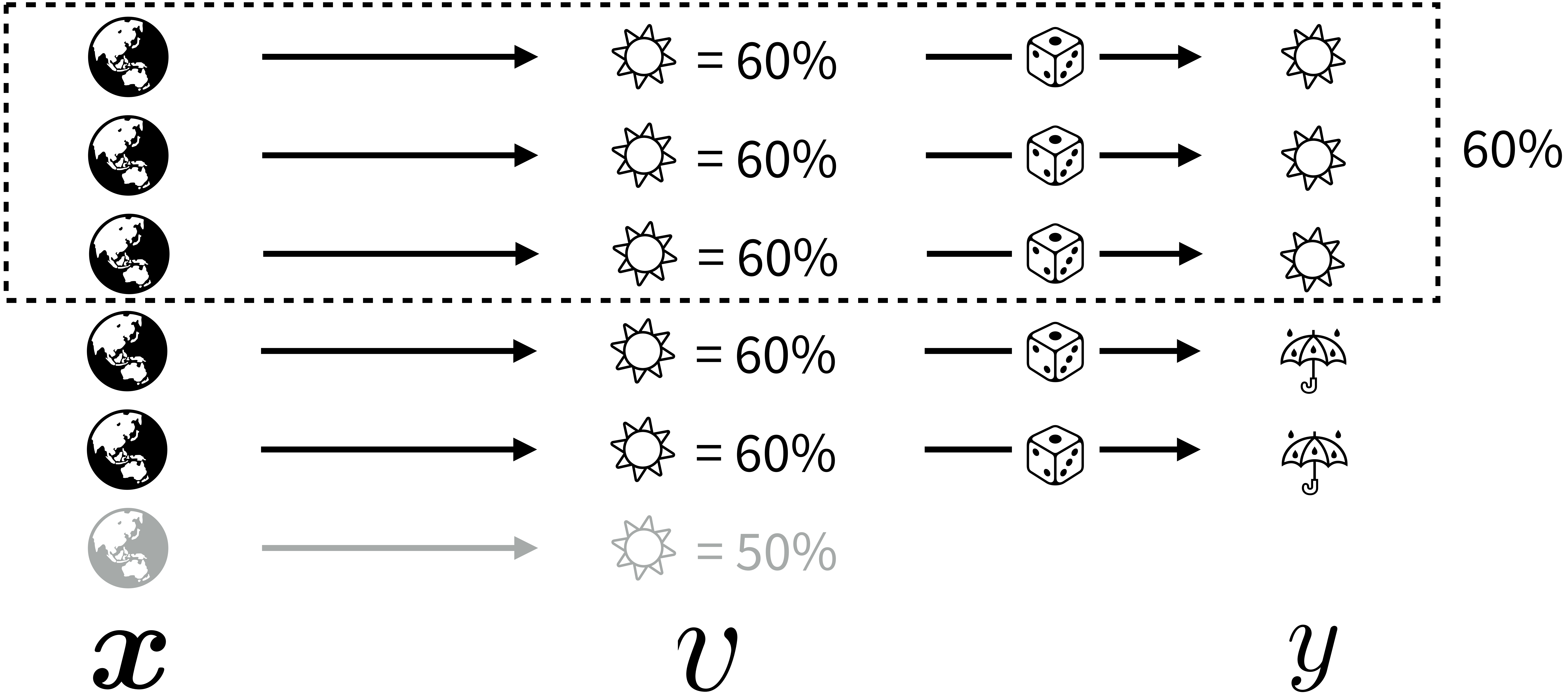
Classification with probabilistic prediction



Q. How reasonable is $\widehat{p}$?

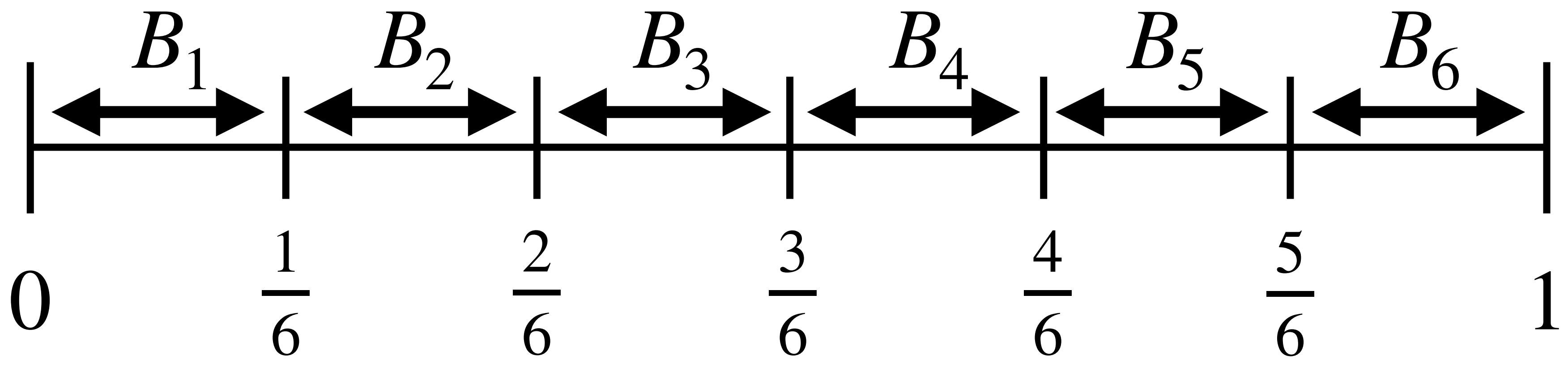
Calibrated prediction

- Predictor $f : \mathcal{X} \to [0,1]$ is **calibrated** iff $\mathbb{E}[Y = 1 | f(X) = v] = v$ holds for any $v \in [0,1]$



Calibrated prediction

- Predictor $f:\mathcal{X}\to[0,1]$ is **calibrated** iff $\mathbb{E}[Y=1|f(X)=v]=v$ holds for any $v\in[0,1]$
- How to measure calibration?
  - **ECE (Expected Calibraiton Error)**



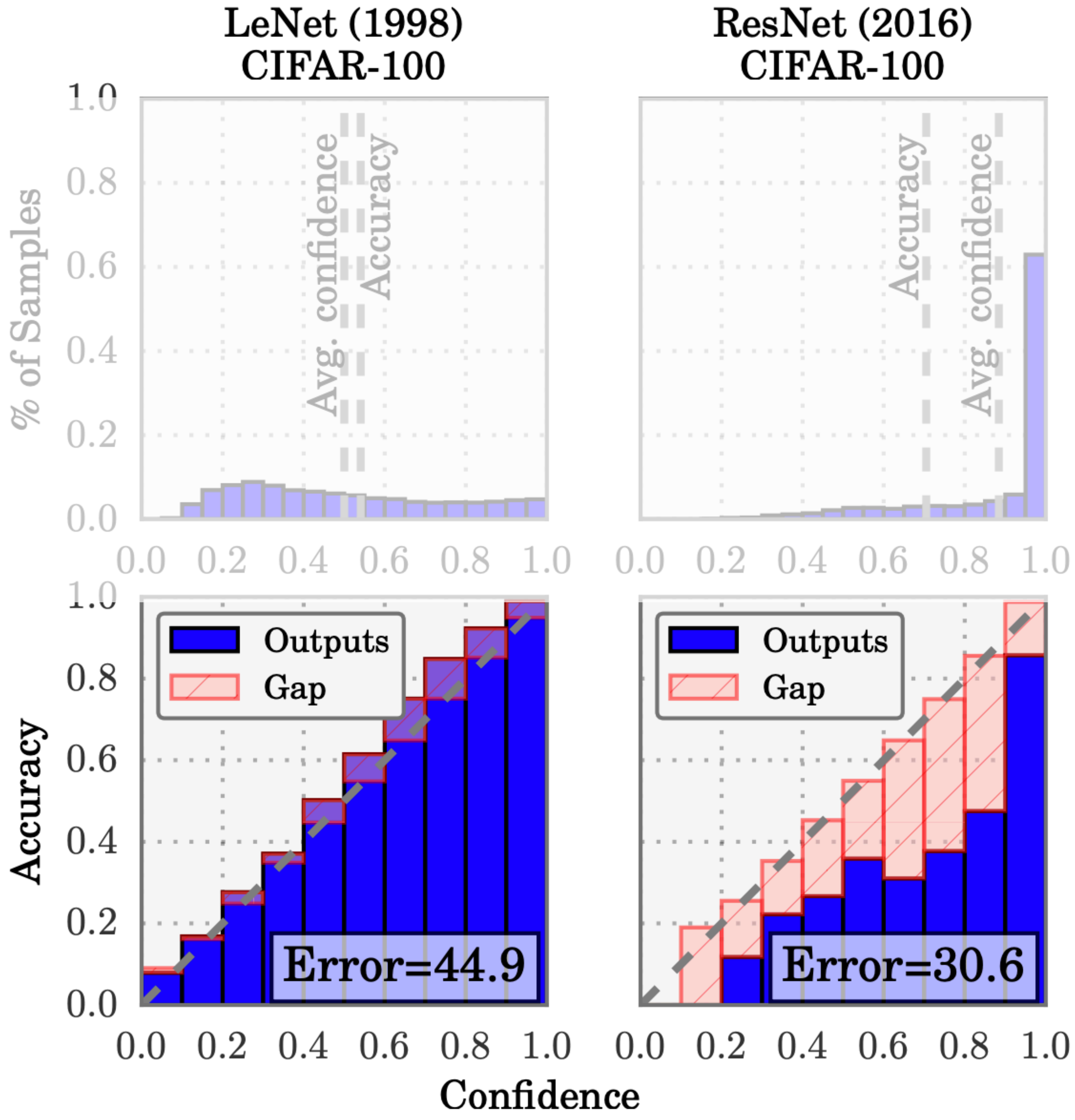
$$\mathrm{ECE}=\sum_{m=1}^{M}\frac{|B_m|}{n}\left|\mathrm{acc}(B_m)-\mathrm{conf}(B_m)\right|$$

fraction of Y=1 samples in bin B_m (corresp. to $\mathbb{E}[Y=1\,|\,f(X)=v]$)

confidence of predictor in bin B_m (corresp. to v)

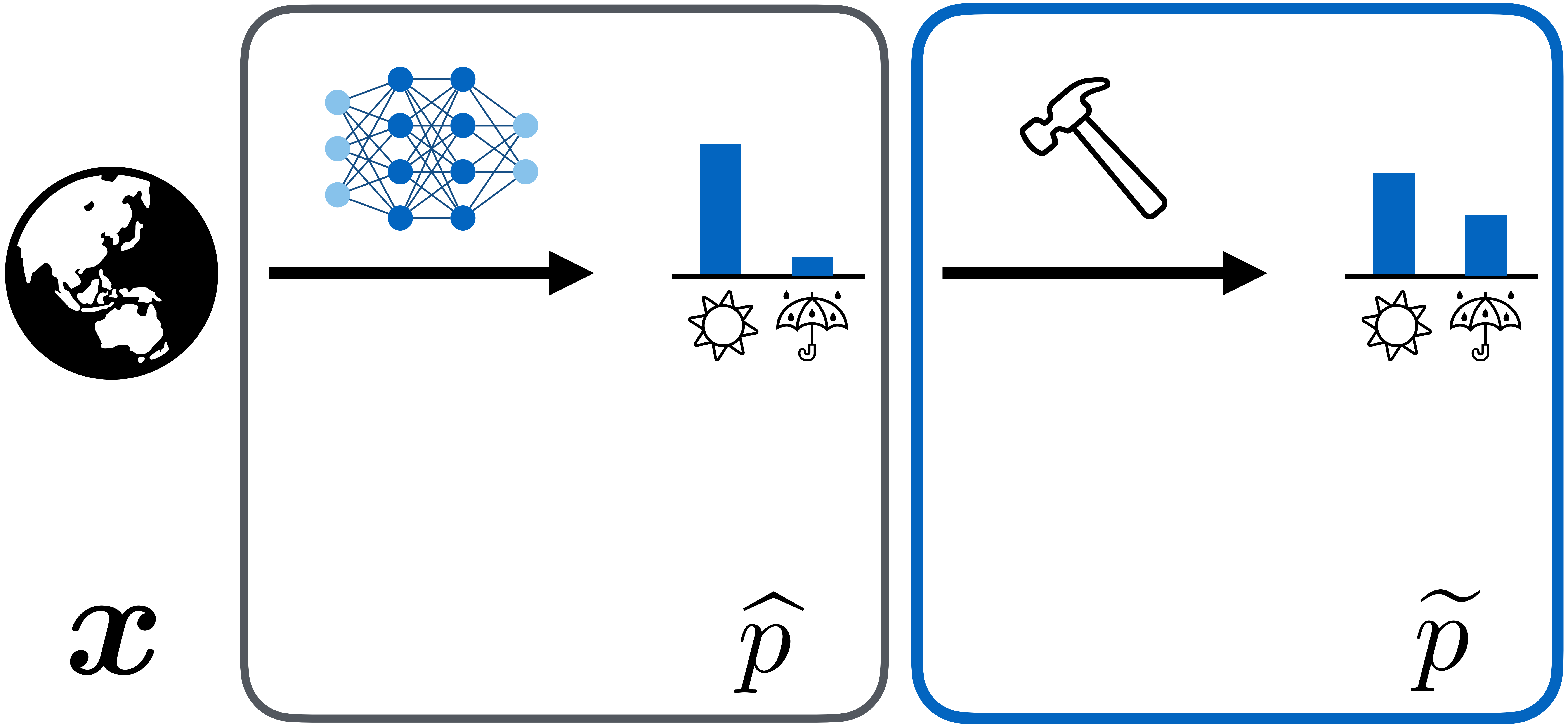
Modern neural nets are not well-calibrated

[Guo+ 2017]



Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *ICML.*

Post-hoc calibration



not well-calibrated

post-hoc calibration (recalibration)

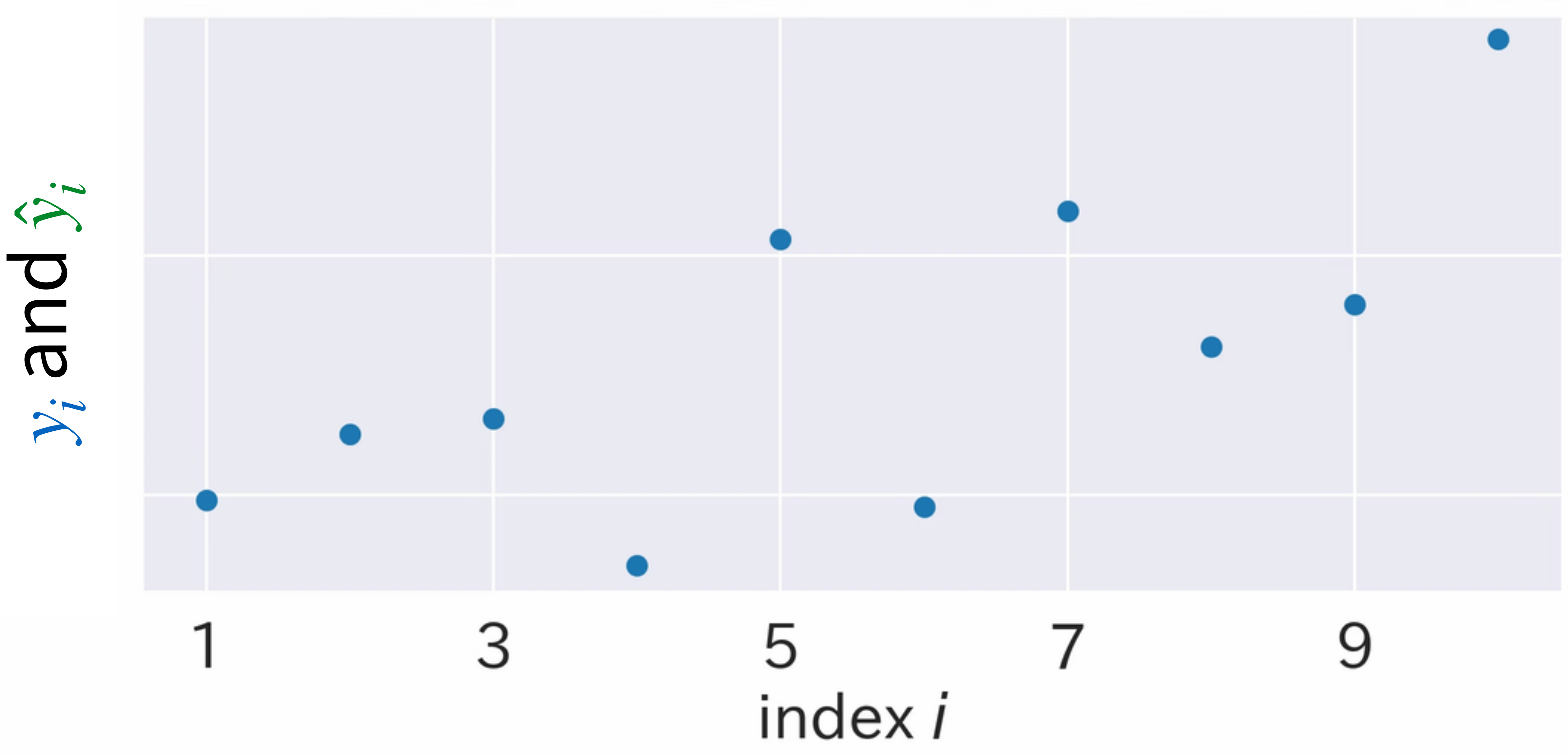
Isotonic regression

$$\min_{\hat{y}_1,\ldots,\hat{y}_n\in\mathbb{R}} \sum_{i\in[n]} (y_i - \hat{y}_i)^2$$

regression error to true label y

$$\text{subject to} \quad (z_i - z_j)(\hat{y}_i - \hat{y}_j) \geq 0 \quad \forall(i,j)\in[n]^2.$$

regression input z and output $\hat{y}$ are monotone

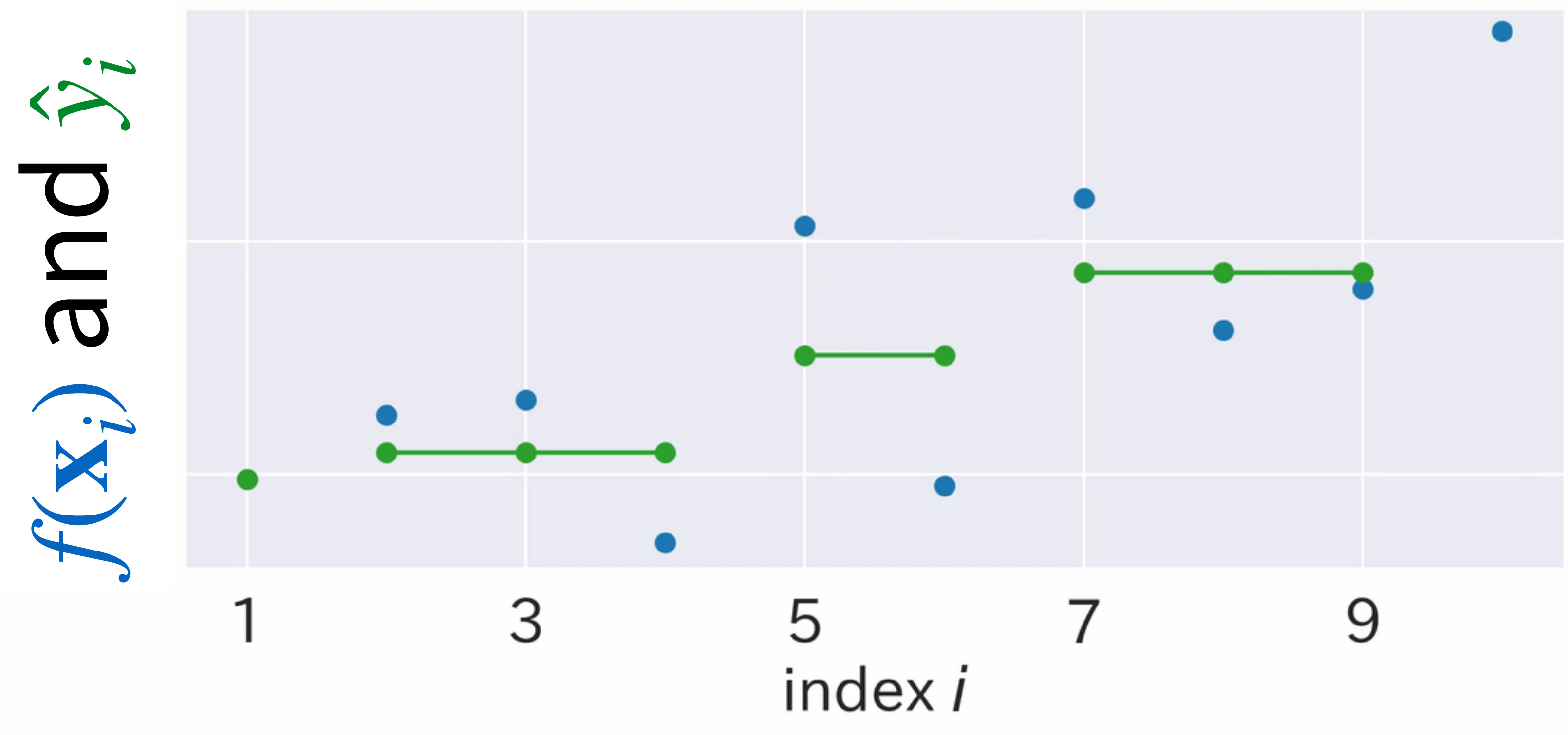


Isotonic regression as recalibrator

$$\min_{\hat{y}_1,\dots,\hat{y}_n\in\mathbb{R}} \sum_{i\in[n]} (y_i - \hat{y}_i)^2$$

$$\text{subject to} \quad \boxed{(f(\mathbf{x}_i) - f(\mathbf{x}_j)})(\hat{y}_i - \hat{y}_j) \geq 0 \quad \forall(i,j) \in [n]^2.$$

use raw prediction $f(x)$ as regression input

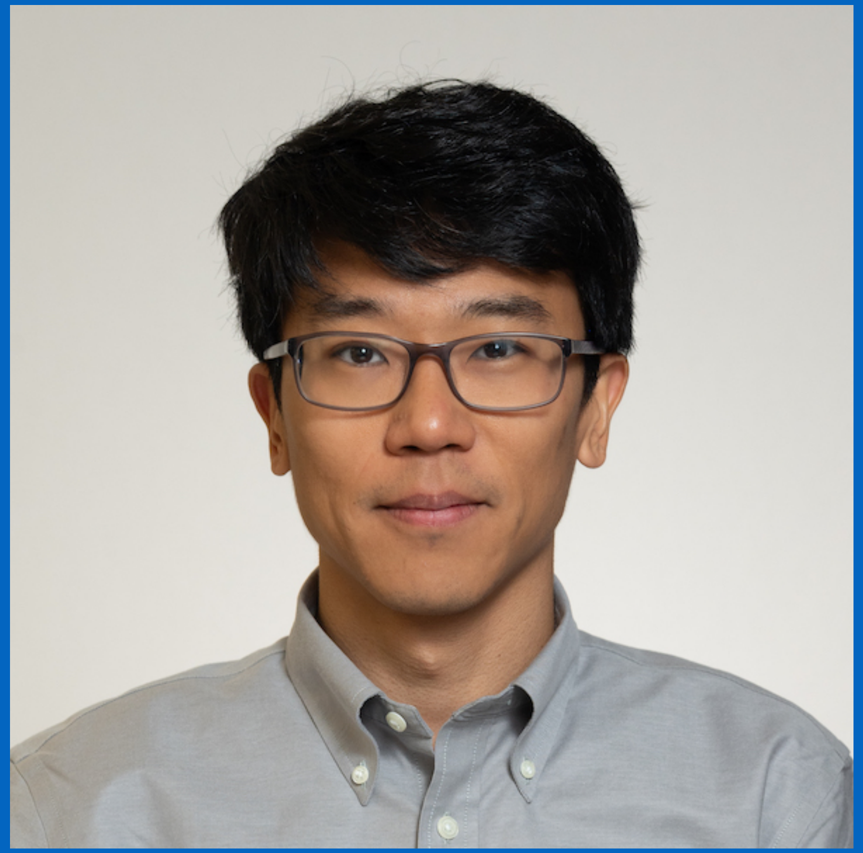


- Good news: IR minimizes ECE (on the training set) [Berta+ 2024]
- Bad news: IR is only **applicable to binary classification**
  - Why? Because "monotonicity" cannot be straightforwardly extended beyond $\mathbb{R}$

E. Berta, F. Bach, and M. Jordan. (2024) Classifier calibration with ROC-regularized isotonic regression. In *AISTATS*.

Brenier Isotonic Regression

Joint work with Amirreza and Yutong, and will be presented at AISTATS2026



Multivariate monotonicity is non-trivial

$$\min_{\hat{\mathbf{y}}_1,\ldots,\hat{\mathbf{y}}_n \in \mathbb{R}^K} \sum_{i\in[n]} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2$$

$$\text{subject to } \mathbf{f}(\mathbf{x}_i) \in \mathbb{R}^K \text{ and } \hat{\mathbf{y}}_i \in \mathbb{R}^K \text{ are ``monotone''}$$

???

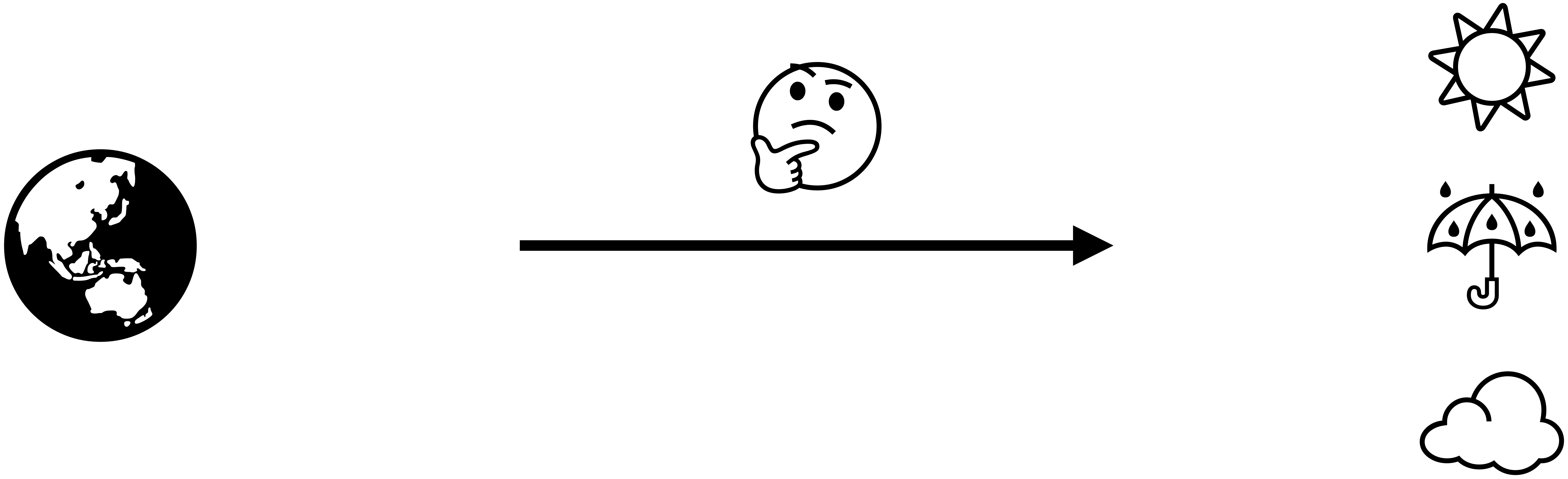
- Coordinate-wise monotonicity?

$$\mathbf{z} \preceq \mathbf{z}' \iff z_i \leq z_i' \text{ for } \forall i \in [K]$$

- Operator monotonicity?

$$\langle \mathbf{f}(\mathbf{x}_i) - \mathbf{f}(\mathbf{x}_j), \hat{\mathbf{y}}_i - \hat{\mathbf{y}}_j \rangle \geq 0 \text{ for } \forall i, j \in [n]$$

Multiclass classification



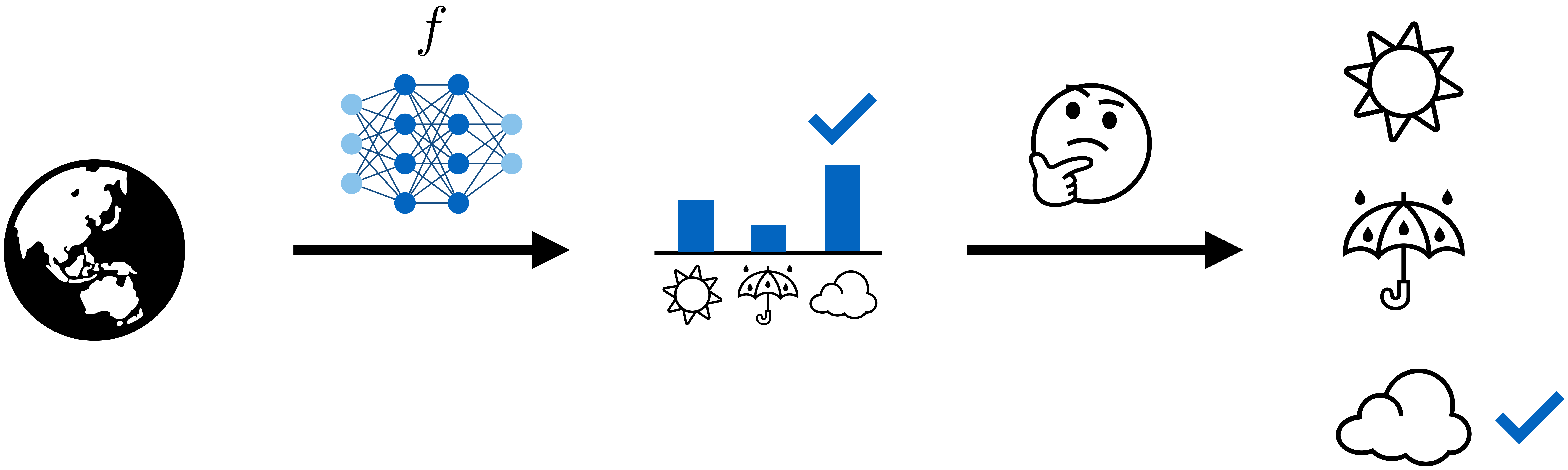
x

feature

$\widehat{y}$

class

Multiclass classification



$\boldsymbol{x}$

feature

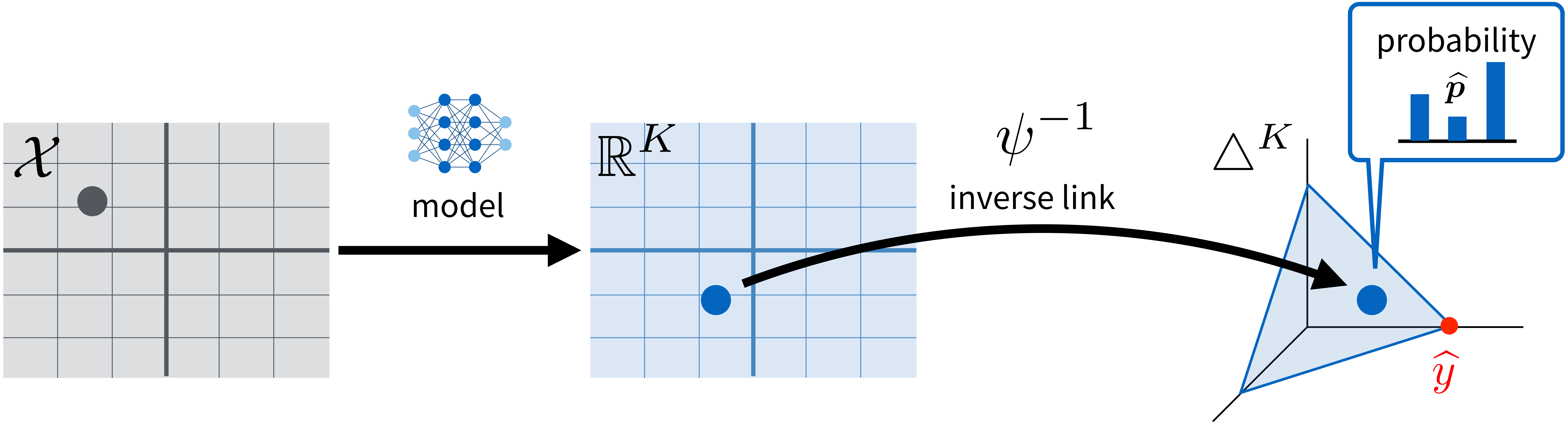
$\boldsymbol{r} = f(\boldsymbol{x})$

report

$\widehat{y}$

class

Multiclass classification in GLM viewpoint



model

inverse link

probability

$\boldsymbol{x}$

feature

$\boldsymbol{r} = f(\boldsymbol{x})$

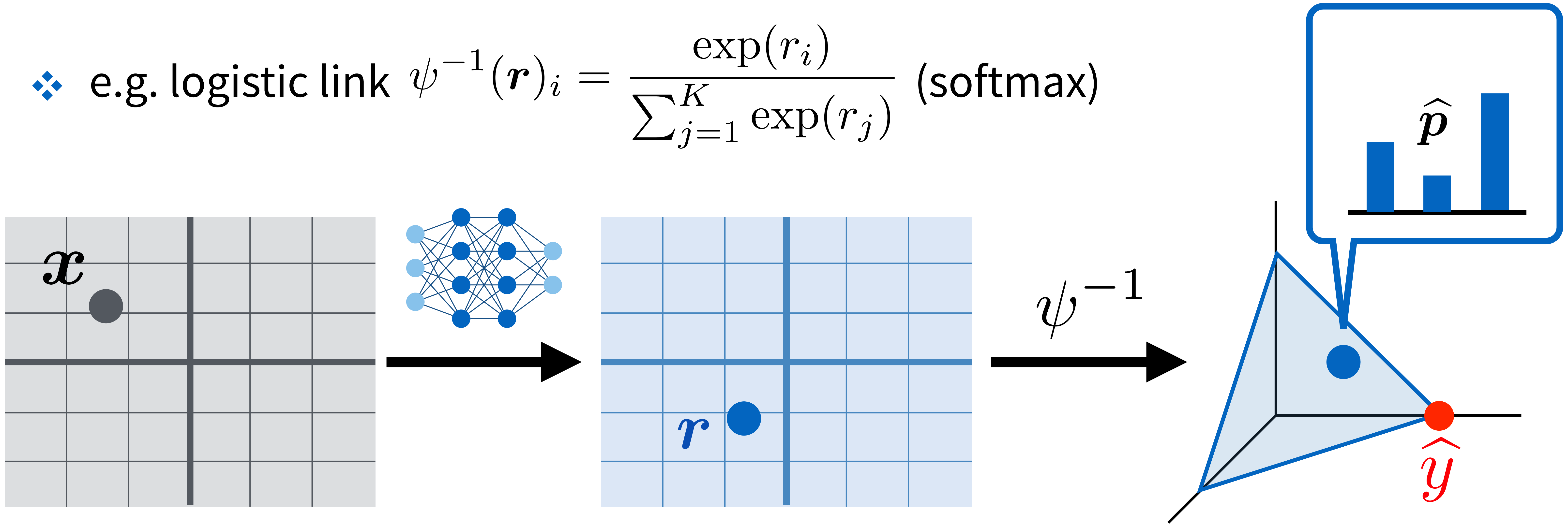
report

$\widehat{y}$

class

Multiclass classification in GLM viewpoint

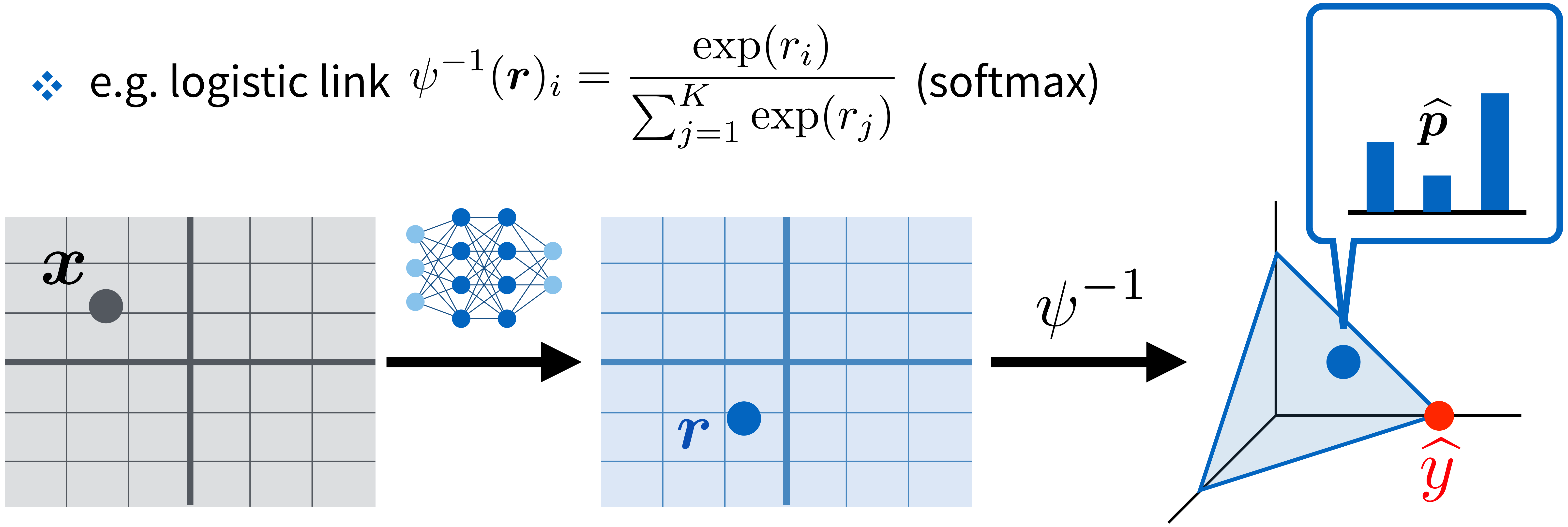
- Define report space $\mathbb{R}^K$ for K-class classification
- Inverse link function ψ^{-1} maps report $\boldsymbol{r} \in \mathbb{R}^K$ to prediction $\widehat{\boldsymbol{p}} \in \triangle^K$
  - e.g. logistic link $\psi^{-1}(\boldsymbol{r})_i = \frac{\exp(r_i)}{\sum_{j=1}^K \exp(r_j)}$ (softmax)



$$\widehat{\boldsymbol{p}} = \mathbb{E}[Y|X = \boldsymbol{x}] = \psi^{-1}(\boldsymbol{r}) = \psi^{-1}(\boldsymbol{W}^*\boldsymbol{x})$$

Multiclass classification in GLM viewpoint

- Define report space $\mathbb{R}^K$ for K-class classification
- Inverse link function ψ^{-1} maps report $\boldsymbol{r} \in \mathbb{R}^K$ to prediction $\widehat{\boldsymbol{p}} \in \triangle^K$
  - e.g. logistic link $\psi^{-1}(\boldsymbol{r})_i = \dfrac{\exp(r_i)}{\sum_{j=1}^K \exp(r_j)}$ (softmax)

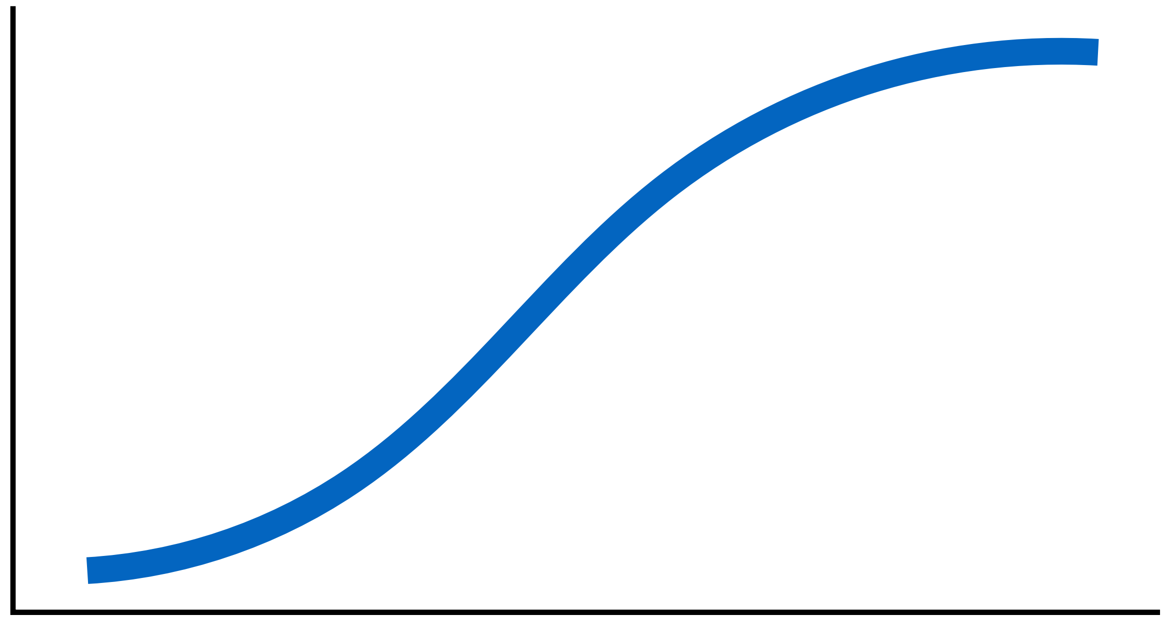


$$\widehat{\boldsymbol{p}} = \mathbb{E}[Y | X = \boldsymbol{x}] = \nabla\Phi(\boldsymbol{W}^*\boldsymbol{x})$$

gradient of **convex** potential Φ

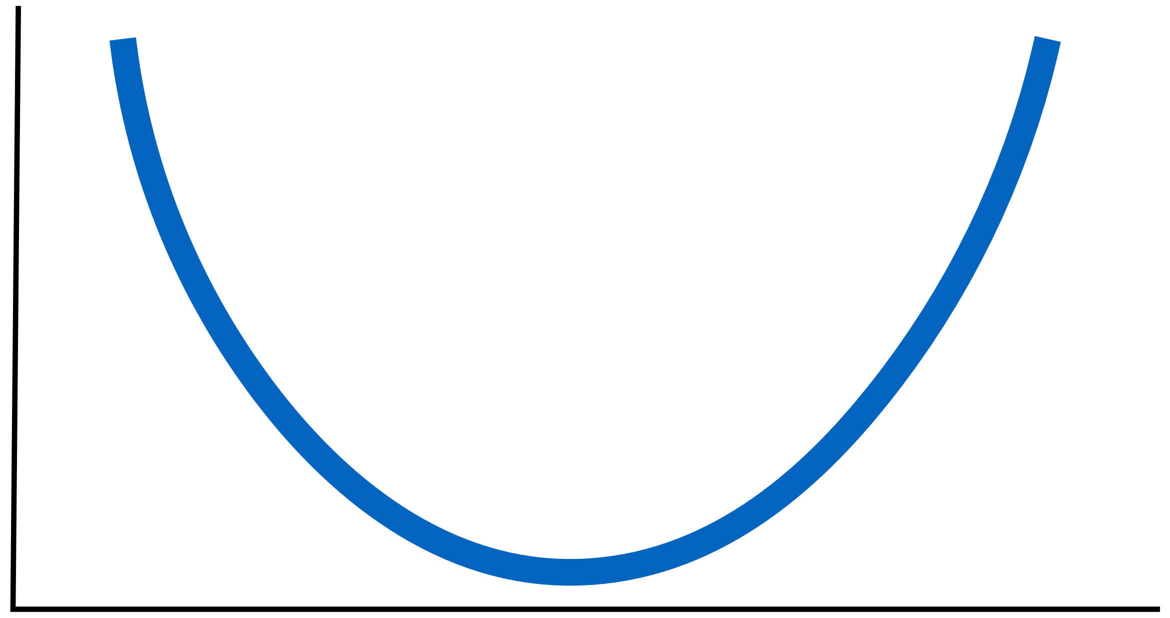
univariate

monotone function



$\Big\uparrow \dfrac{\mathrm{d}}{\mathrm{d}x}$

convex function

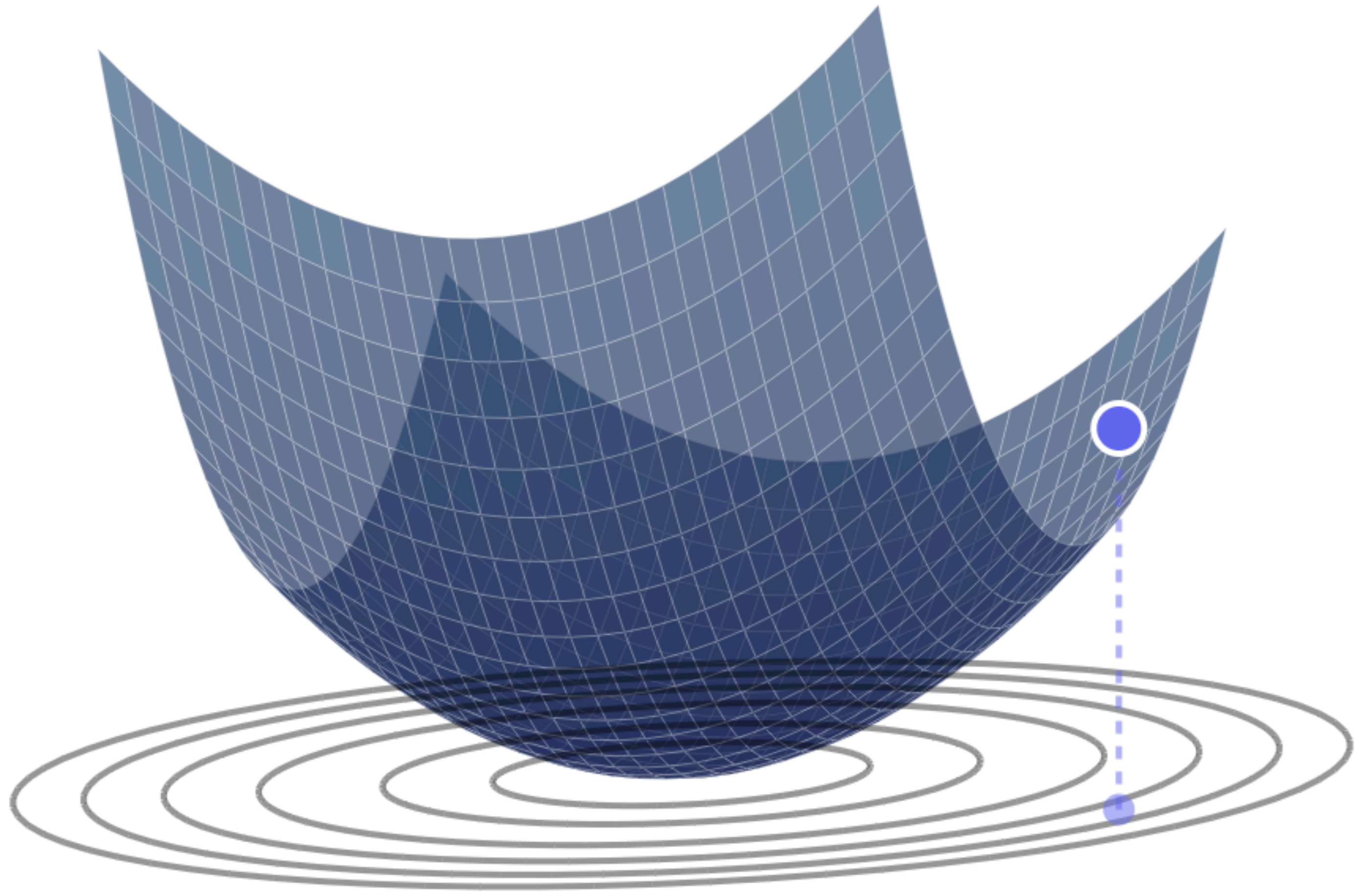


$\longrightarrow$

multivariate

cyclic monotone function

$\Big\uparrow \nabla$



Cyclic monotonicity

[Rockafellar 1966]

Definition A relation $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$ is cyclic monotone if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[m]} \subseteq \Gamma$, the following inequality holds (with convention $\mathbf{y}_{m+1} = \mathbf{y}_1$):

$$\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \leq \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$

- Keep in mind:

$$\{(\boxed{\mathbf{x}_i}, \boxed{\mathbf{y}_i})\}_{i\in[m]} \subseteq \Gamma$$

function input

corresponding gradient “$\nabla\Phi(\mathbf{x}_i)$”

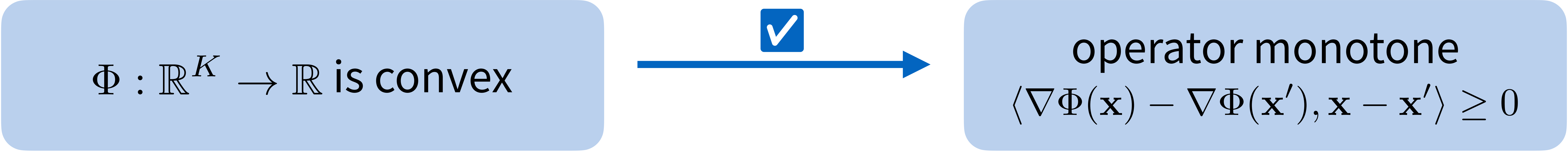
R. T. Rockafellar. Characterization of the subdifferentials of convex functions. Pacific Journal of Mathematics, 17(3):497–510, 1966.

Convexity and monotonicity

[Rockafellar 1966]

Definition A relation $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$ is cyclic monotone if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[m]} \subseteq \Gamma$, the following inequality holds (with convention $\mathbf{y}_{m+1} = \mathbf{y}_1$):

$$\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \leq \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$



$\Phi : \mathbb{R}^K \to \mathbb{R}$ is convex

☑

operator monotone

$\langle \nabla\Phi(\mathbf{x}) - \nabla\Phi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq 0$

R. T. Rockafellar. Characterization of the subdifferentials of convex functions. Pacific Journal of Mathematics, 17(3):497–510, 1966.

Convexity and monotonicity

[Rockafellar 1966]

Definition A relation $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$ is cyclic monotone if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[m]} \subseteq \Gamma$, the following inequality holds (with convention $\mathbf{y}_{m+1} = \mathbf{y}_1$):

$$\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \leq \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$

there exists convex Φ s.t. $\nabla\Phi = \varphi$

☑ →

operator monotone $\langle \varphi(\mathbf{x}) - \varphi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq 0$

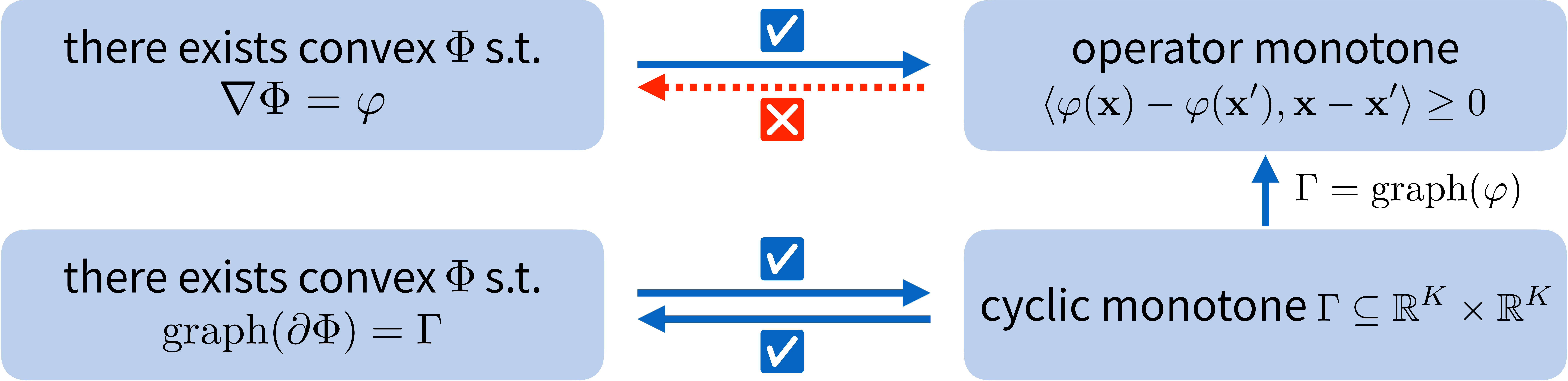
R. T. Rockafellar. Characterization of the subdifferentials of convex functions. Pacific Journal of Mathematics, 17(3):497–510, 1966.

Convexity and monotonicity

[Rockafellar 1966]

Definition A relation $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$ is cyclic monotone if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [m]} \subseteq \Gamma$, the following inequality holds (with convention $\mathbf{y}_{m+1} = \mathbf{y}_1$):

$$\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \leq \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$



there exists convex Φ s.t. $\nabla \Phi = \varphi$

☑

☒

operator monotone $\langle \varphi(\mathbf{x}) - \varphi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq 0$

$\Gamma = \mathrm{graph}(\varphi)$

there exists convex Φ s.t. $\mathrm{graph}(\partial \Phi) = \Gamma$

☑

☑

cyclic monotone $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$

R. T. Rockafellar. Characterization of the subdifferentials of convex functions. Pacific Journal of Mathematics, 17(3):497–510, 1966.

“Cyclic monotone” isotonic regression

Definition A relation $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$ is cyclic monotone if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[m]} \subseteq \Gamma$, the following inequality holds (with convention $\mathbf{y}_{m+1} = \mathbf{y}_1$):

$$\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \leq \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$

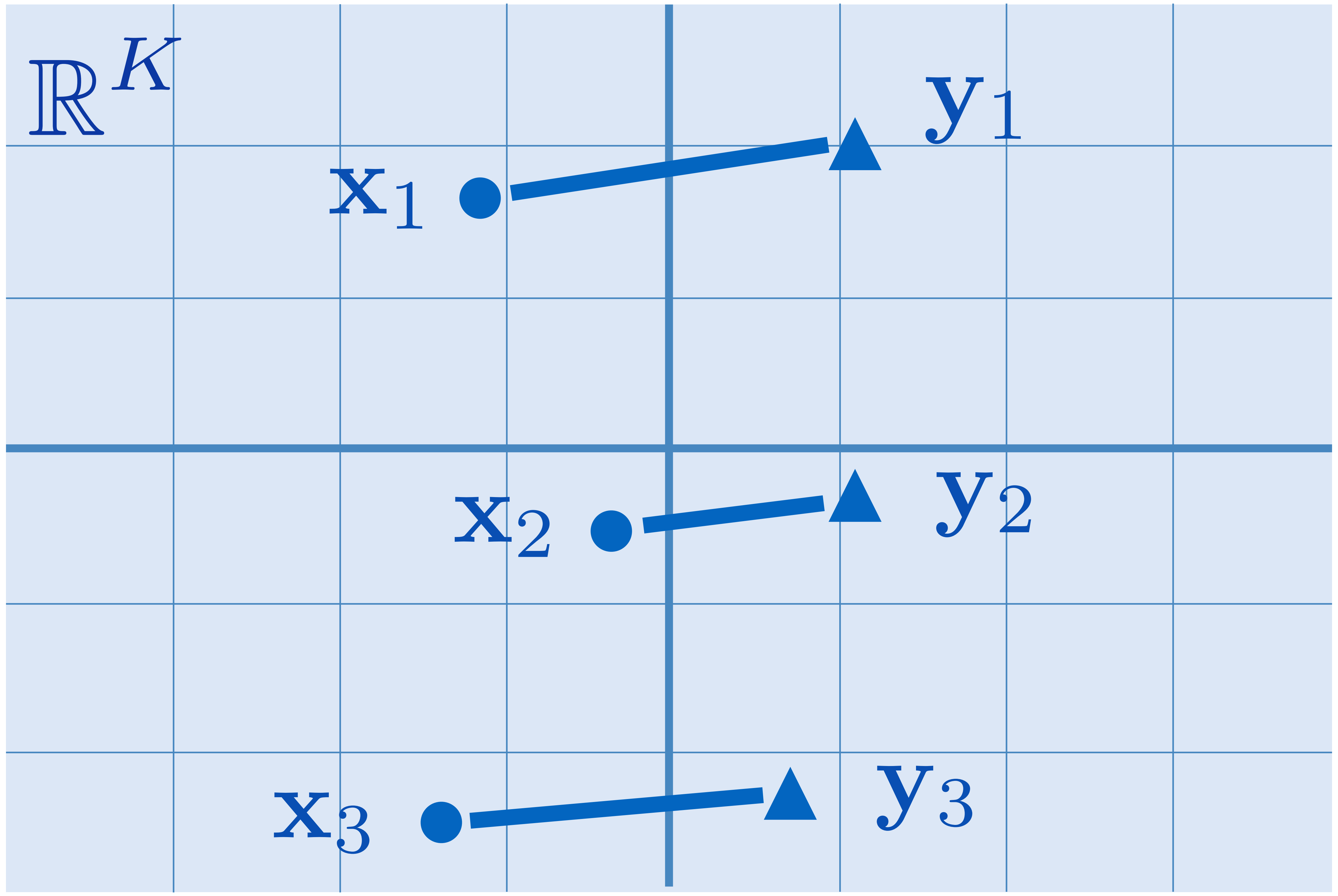
$$\min_{\hat{\mathbf{y}}_1,\dots,\hat{\mathbf{y}}_n \in \mathbb{R}^K} \sum_{i\in[n]} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2$$

$$\text{subject to } \{(\mathbf{f}(\mathbf{x}_i), \hat{\mathbf{y}}_i)\}_{i\in[n]} \text{ is cyclic monotone}$$

Q. How to impose this?

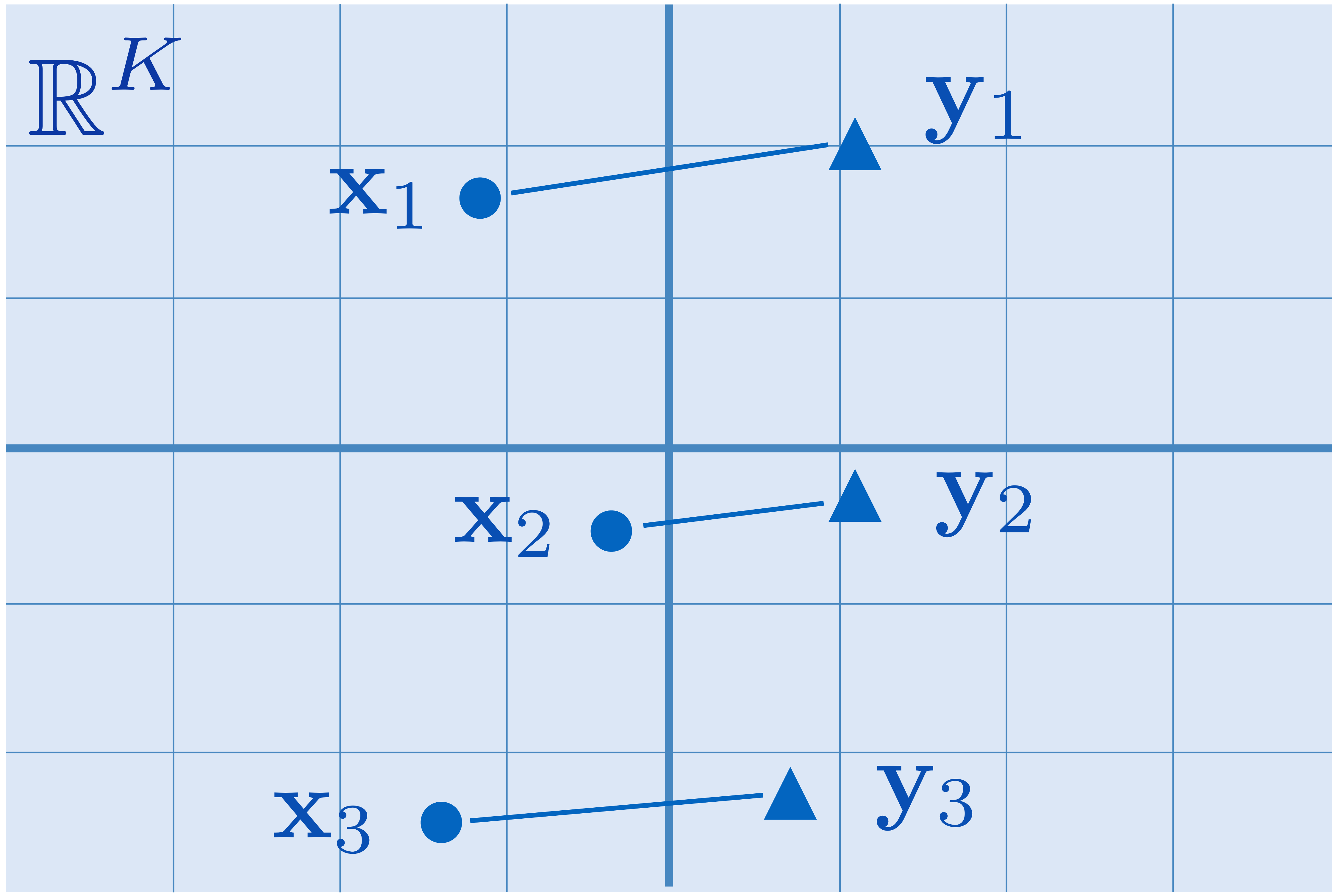
Definition A relation $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$ is cyclic monotone if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[m]} \subseteq \Gamma$, the following inequality holds (with convention $\mathbf{y}_{m+1} = \mathbf{y}_1$):

$$\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \leq \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$



Definition A relation $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$ is cyclic monotone if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[m]} \subseteq \Gamma$, the following inequality holds (with convention $\mathbf{y}_{m+1} = \mathbf{y}_1$):

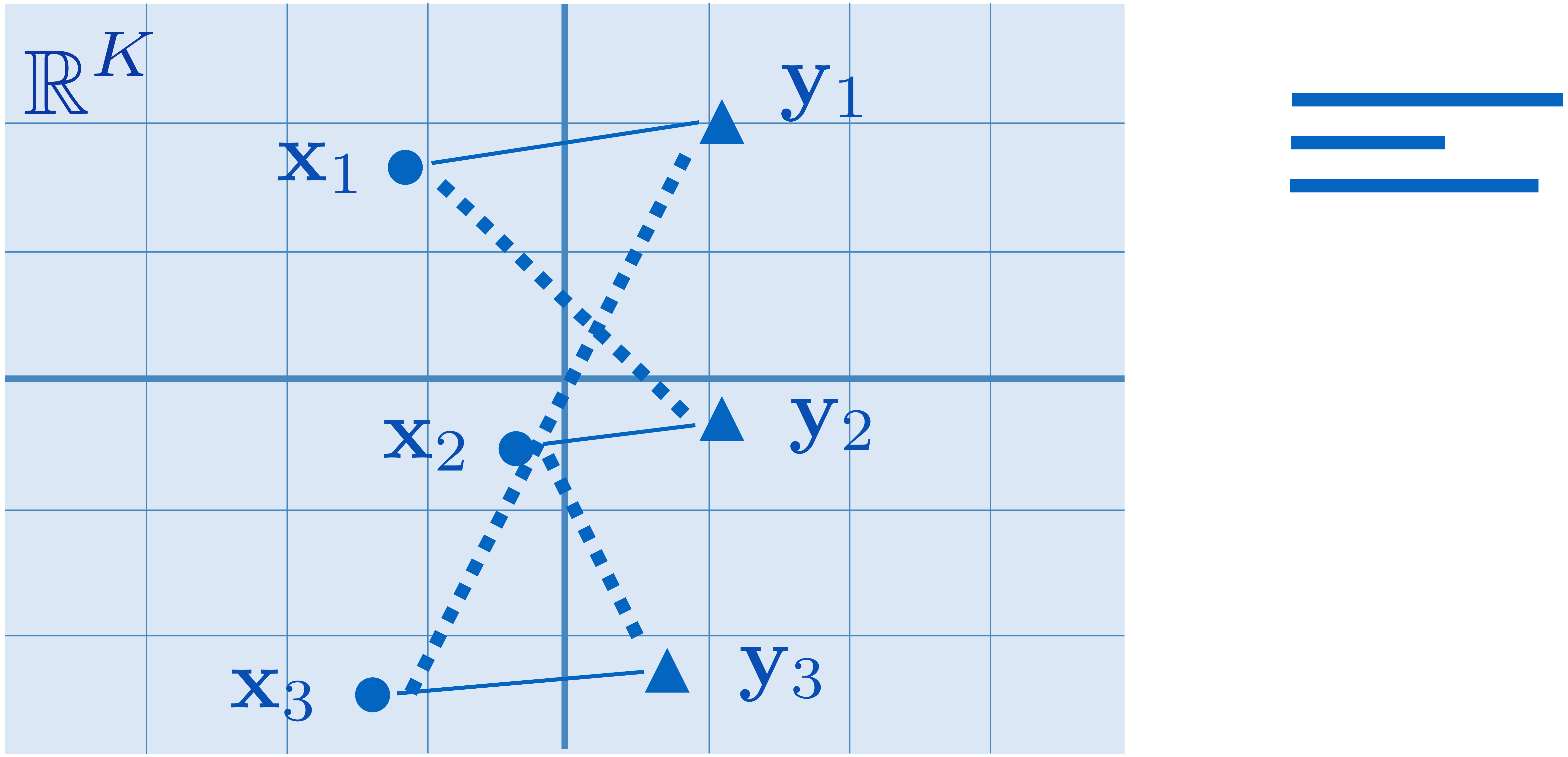
$$\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \leq \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$



$$= \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2$$

Definition A relation $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$ is cyclic monotone if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[m]} \subseteq \Gamma$, the following inequality holds (with convention $\mathbf{y}_{m+1} = \mathbf{y}_1$):

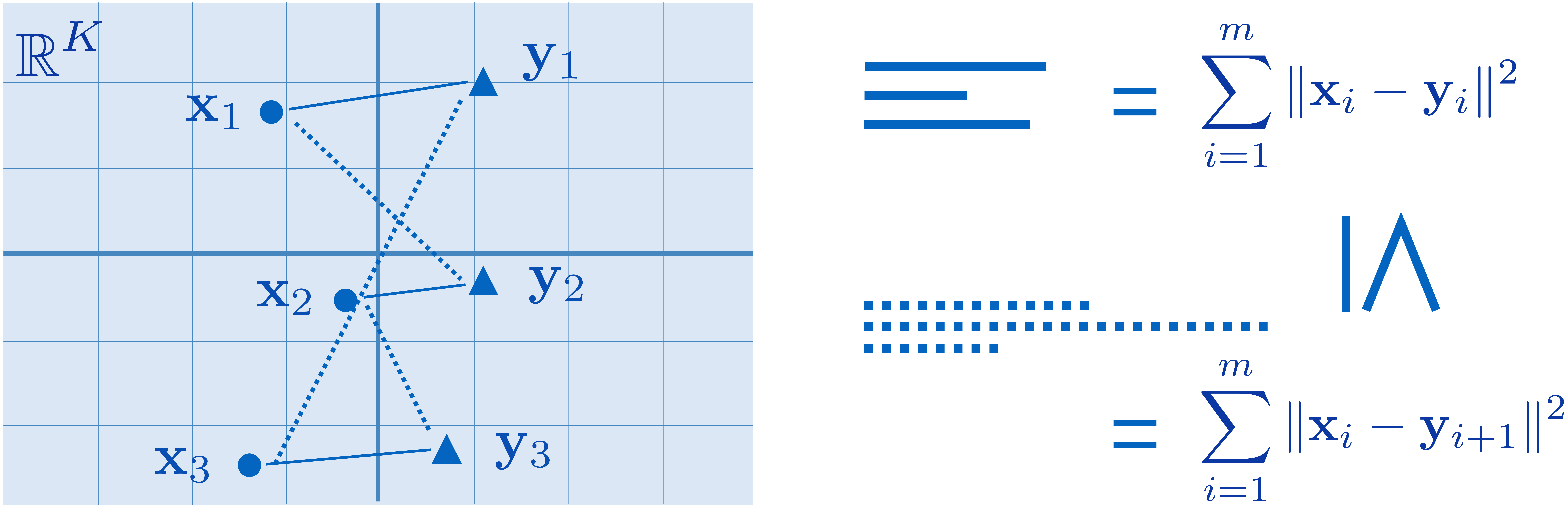
$$\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \leq \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$



$$= \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2$$

Definition A relation $\Gamma \subseteq \mathbb{R}^K \times \mathbb{R}^K$ is cyclic monotone if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [m]} \subseteq \Gamma$, the following inequality holds (with convention $\mathbf{y}_{m+1} = \mathbf{y}_1$):

$$\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2 \leq \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$



$$= \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_i\|^2$$

$$\rotatebox{90}{$\leq$}$$

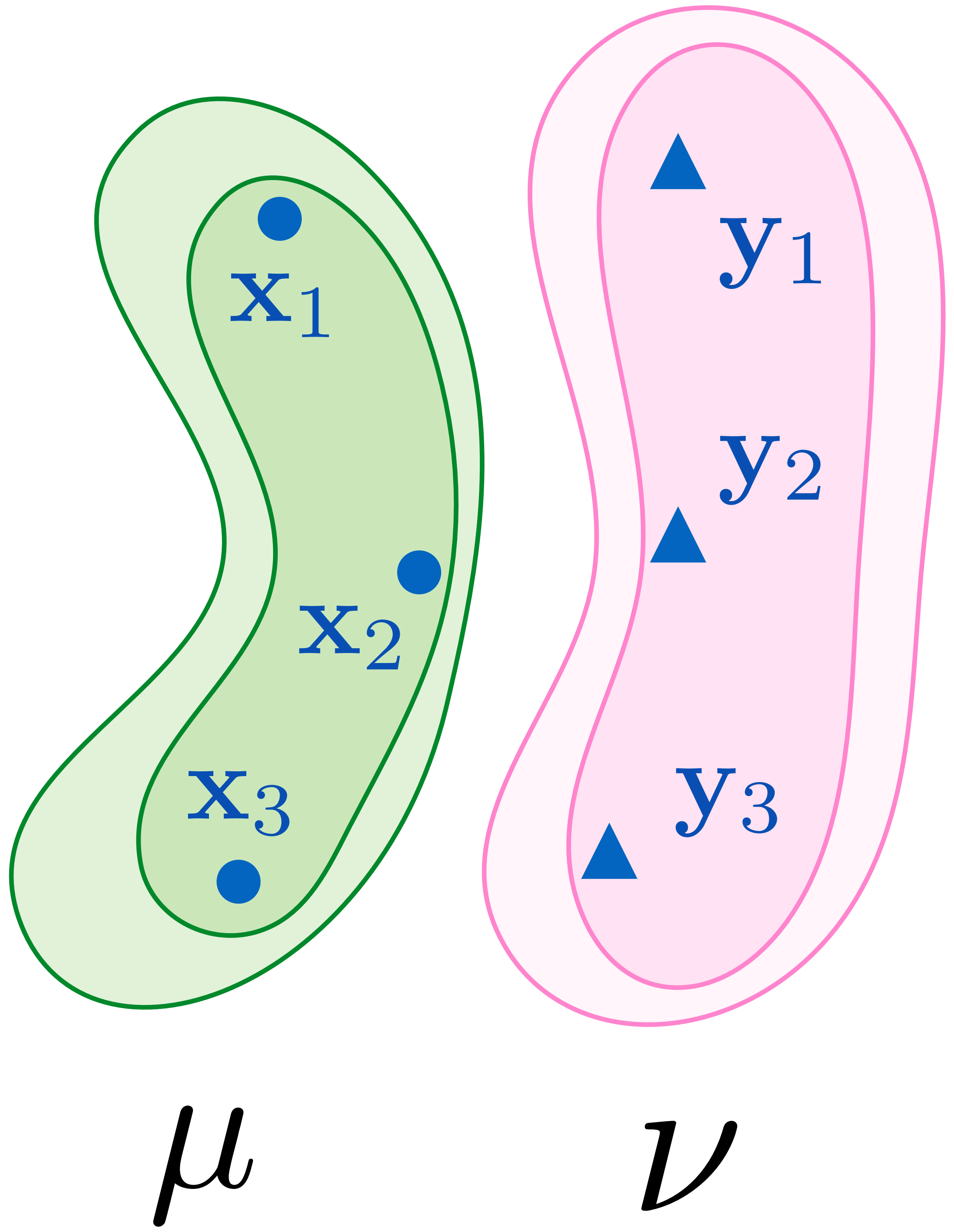
$$= \sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{y}_{i+1}\|^2$$

Optimal transport is cyclic monotone!

- Consider optimal transport problem
  - source $\mu = \frac{1}{n}\sum_{i\in[n]}\delta_{\mathbf{x}_i}$, target $\nu = \frac{1}{n}\sum_{i\in[n]}\delta_{\mathbf{y}_i}$
  - cost $C_{ij} = \|\mathbf{x}_i - \mathbf{y}_j\|^2$

$$\min_{P\in\mathcal{B}(n,n)} \sum_{i,j=1}^{n} C_{ij}P_{ij}$$

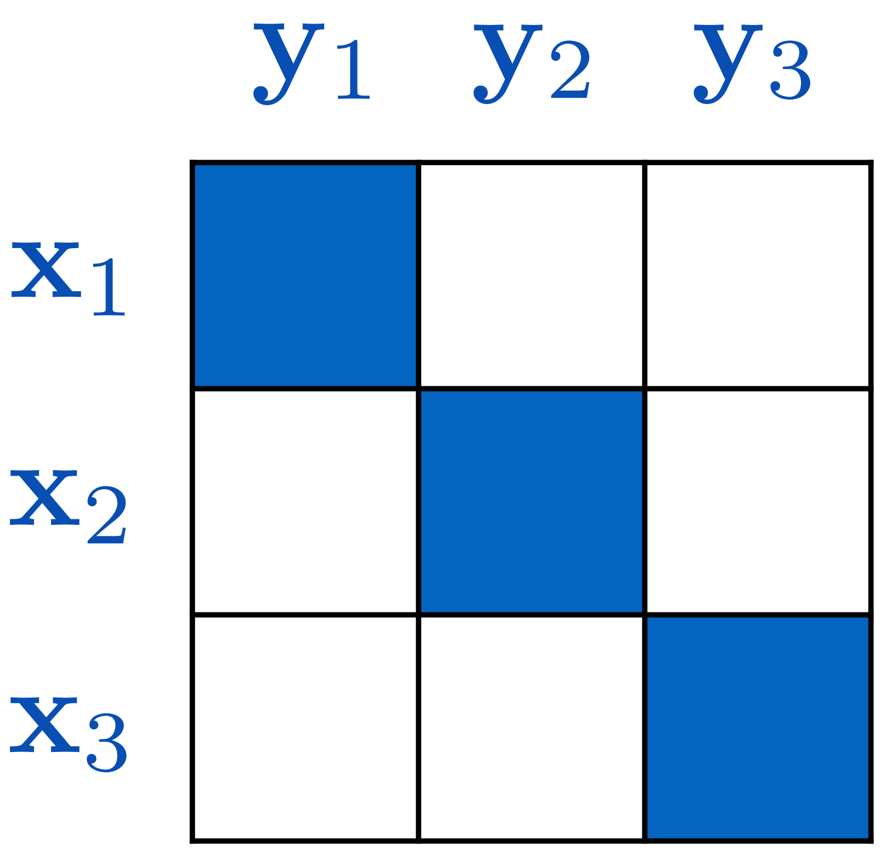
where $\mathcal{B}(n,n) = \{P \in \mathbb{R}^{n\times n} : nP\mathbf{1} = \mathbf{1},\, nP^\top\mathbf{1} = \mathbf{1}\}$ (Birkhoff)



Theorem The optimal coupling P_* has a cyclic monotone support $\mathrm{supp}(P_*)$

Optimal transport is cyclic monotone!

- Consider optimal transport problem
  - source $\mu = \frac{1}{n}\sum_{i\in[n]} \delta_{\mathbf{x}_i}$, target $\nu = \frac{1}{n}\sum_{i\in[n]} \delta_{\mathbf{y}_i}$
  - cost $C_{ij} = \|\mathbf{x}_i - \mathbf{y}_j\|^2$



	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
$\mathbf{x}_1$	■		
$\mathbf{x}_2$		■	
$\mathbf{x}_3$			■

$$\min_{P\in\mathcal{B}(n,n)} \sum_{i,j=1}^{n} C_{ij}P_{ij}$$

where $\mathcal{B}(n,n) = \{P \in \mathbb{R}^{n\times n} : nP\mathbf{1} = \mathbf{1},\, nP^\top\mathbf{1} = \mathbf{1}\}$ (Birkhoff)

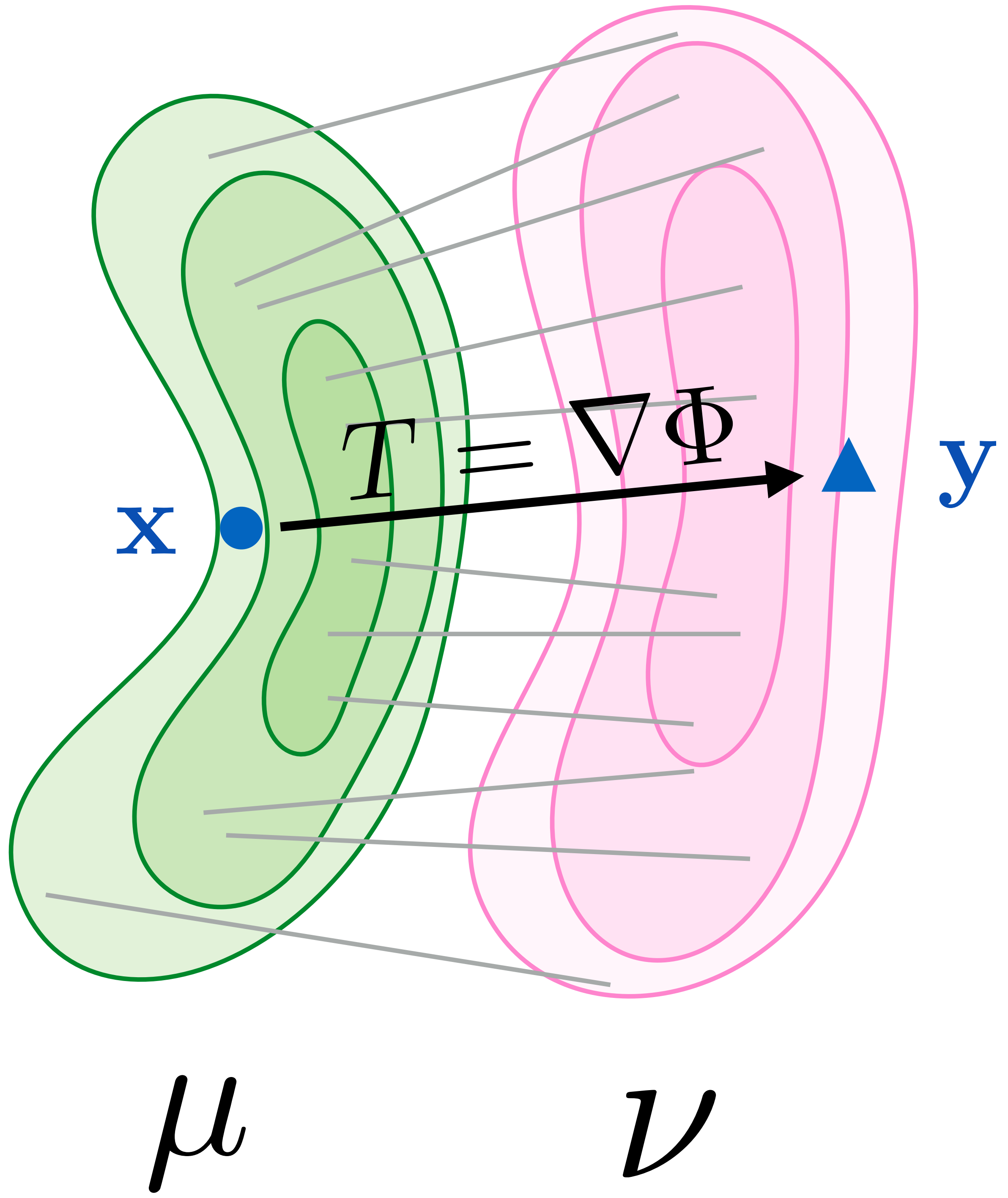
Theorem The optimal coupling P_* has a cyclic monotone support $\mathrm{supp}(P_*)$

More generally …

- Consider optimal transport problem
  - source and target $\mu, \nu \in \mathcal{P}(\mathbb{R}^K)$
  - cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$

$$\inf_{\pi} \left\{ \int c \, \mathrm{d}\pi : \pi \in \mathcal{U}(\pi, \nu) \right\}$$

where $\mathcal{U}(\mu, \nu) = \{\pi : \Pi_{1\sharp}\pi = \mu, \Pi_{2\sharp}\pi = \nu\}$



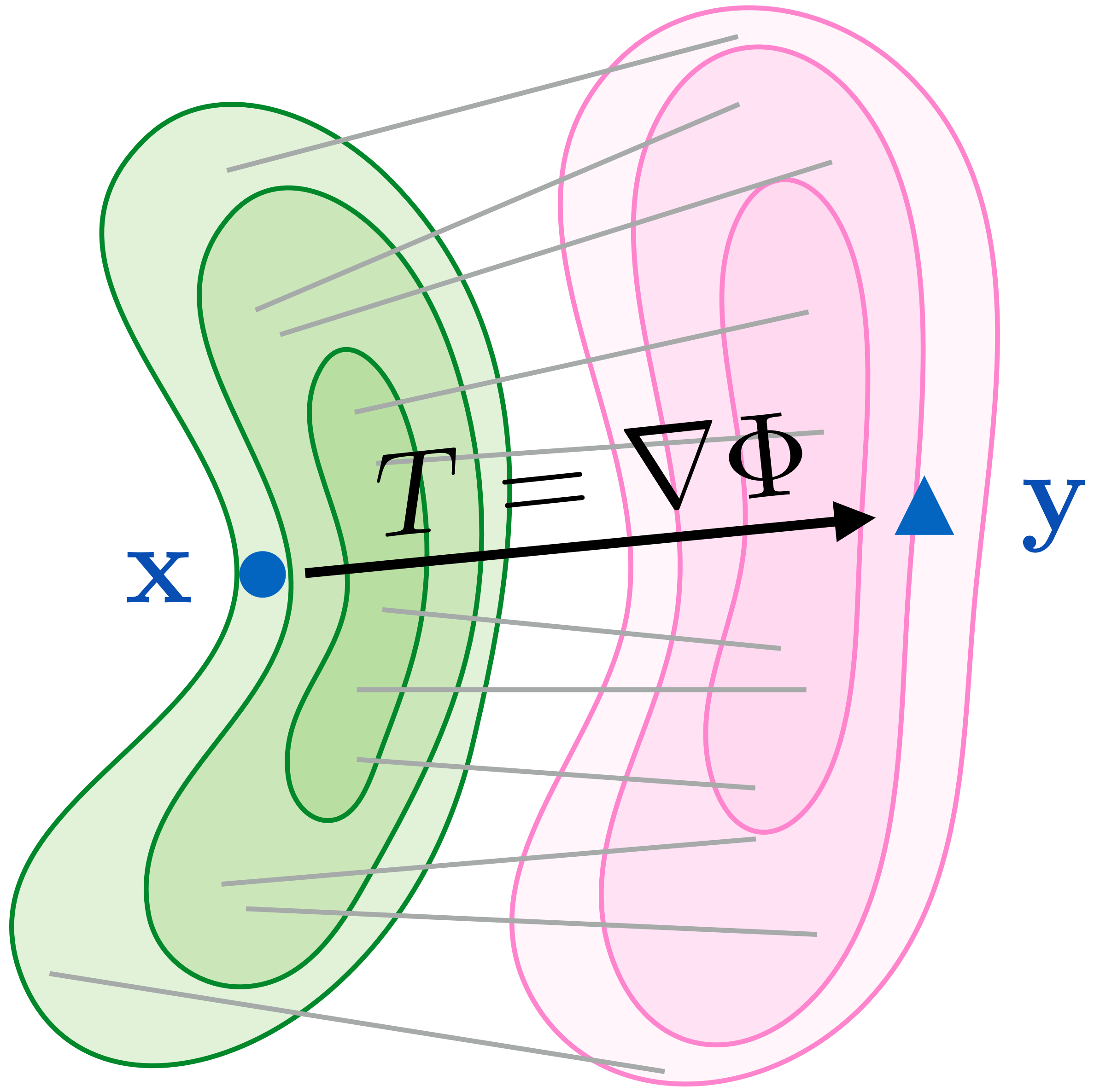
Theorem If μ has a density w.r.t. the Lebesgue measure, there exists the optimal transport map $T : \mathbb{R}^K \to \mathbb{R}^K$ that can be written by $T = \nabla\Phi$ with some differentiable convex potential $\Phi : \mathbb{R}^K \to \mathbb{R}$

[Brenier 1991]

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. (1991)

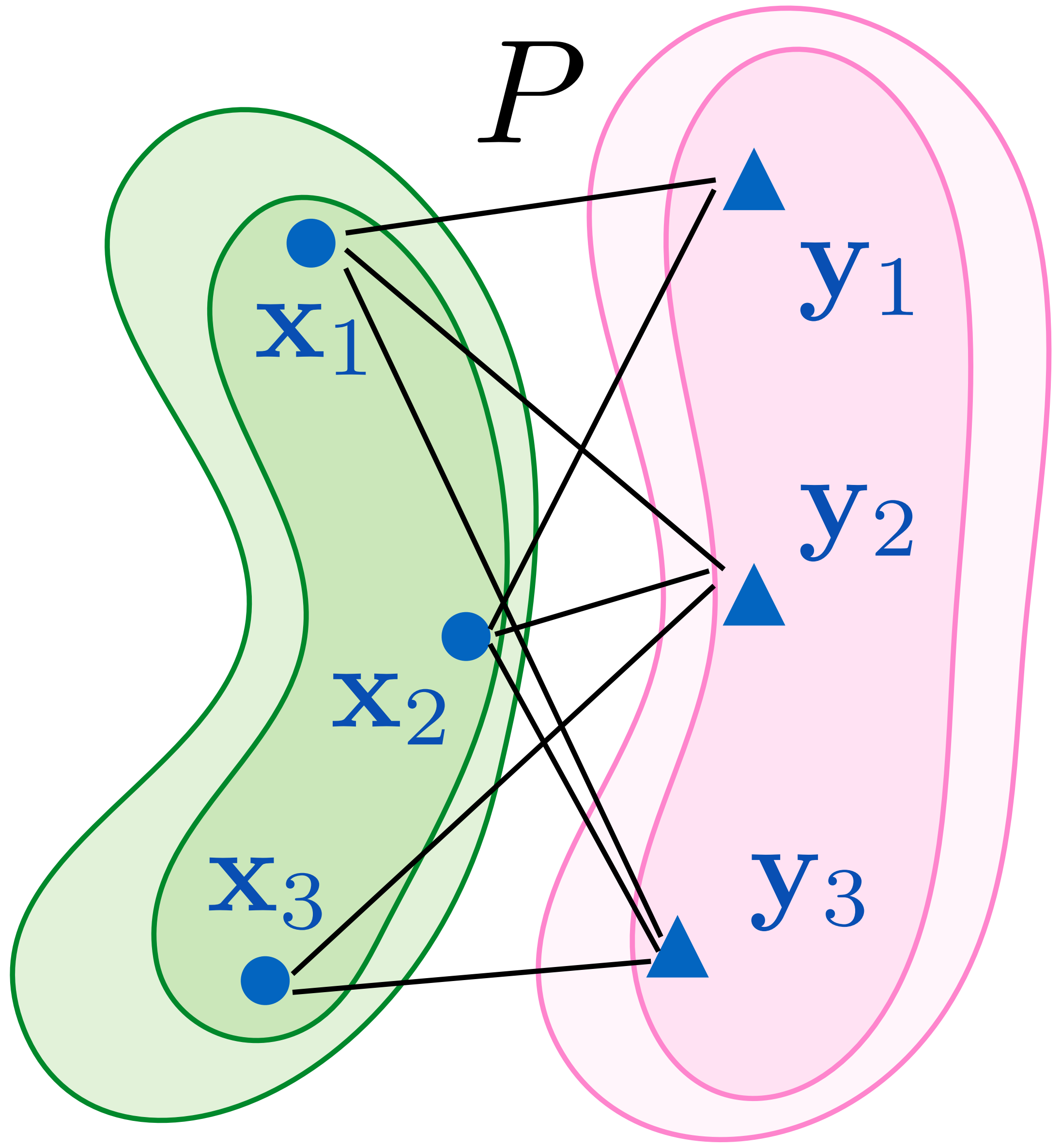
Sum it up | OT & convexity

Regular case



Transport map is $\nabla\Phi$
(Brenier)

General case



Optimal coupling is cyclic monotone
(Kantorovich [Villani 2008, Thm 5.10])

we use this

Cédric Villani. *Optimal Transport: Old and New*, 2008.

Brenier isotonic regression

$$\min_{\hat{\mathbf{y}}_1,\dots,\hat{\mathbf{y}}_n \in \mathbb{R}^K} \sum_{i\in[n]} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2$$

$$\text{subject to } \{(\mathbf{f}(\mathbf{x}_i), \hat{\mathbf{y}}_i)\}_{i\in[n]} \text{ is cyclic monotone}$$

meaning: $\mathbf{f}(\mathbf{x}_i) \mapsto \hat{\mathbf{y}}_j$ is optimal transport map

$$\boxed{\min_{P\in\mathcal{B}(n,n)} \sum_{i,j=1}^{n} C_{ij}P_{ij}} \xrightarrow[P*]{\text{optimal coupling}}$$

OT map = **Barycentric map**

$$T(\mathbf{x}_i) = \frac{\sum_{j\in[n]} P^*_{ij}\mathbf{y}_j}{\sum_{j\in[n]} P^*_{ij}} = n \sum_{j\in[n]} P^*_{ij}\mathbf{y}_j$$

Brenier isotonic regression

$$\min_{\nu} \sum_{i \in [n]} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2$$

$$\text{subject to } \hat{\mathbf{y}}_i = T(\mathbf{f}(\mathbf{x}_i)) \text{ and } T \text{ is OT from } \mu \text{ to } \nu$$

Target $\nu = \frac{1}{n} \sum_{j \in [n]} \delta_{\mathbf{u}_j}$

Barycenteric map (regression func)

$$T(\mathbf{f}(\mathbf{x}_i)) = n \sum_{j \in [n]} P^*_{ij} \mathbf{u}_j$$

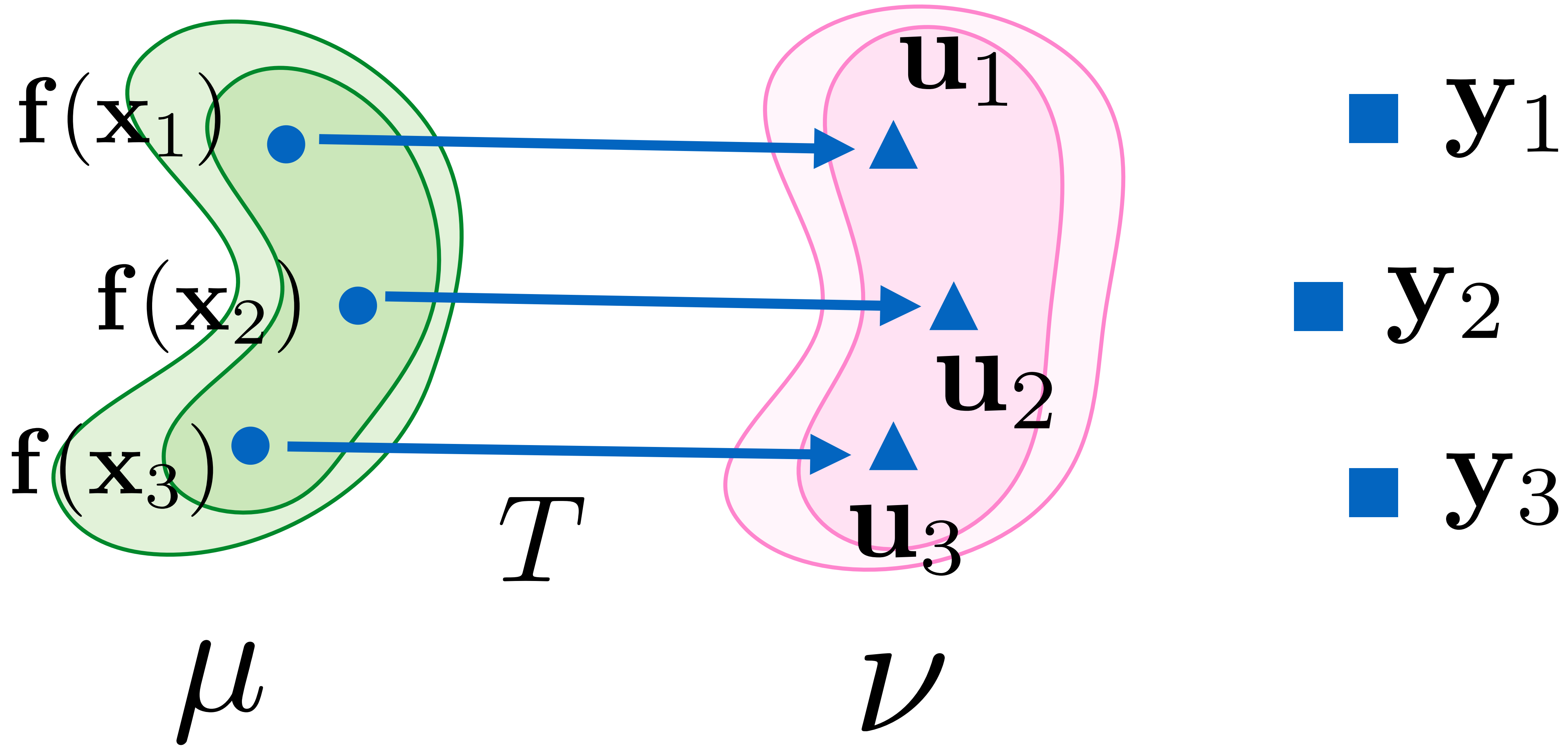
Source $\mu = \frac{1}{n} \sum_{i \in [n]} \delta_{\mathbf{f}(\mathbf{x}_i)}$

Now we minimize regression error with respect to **target measure** ν

Brenier isotonic regression

$$\min_{\mathbf{u}_1,\dots,\mathbf{u}_n} \sum_{i\in[n]} \Bigg\| \mathbf{y}_i - \overbrace{\boxed{n \sum_{j\in[n]} P^*_{ij}\mathbf{u}_j}}^{\text{Barycenteric map } T} \Bigg\|^2$$

$$\text{subject to } P \in \operatorname*{arg\,min}_{P\in\mathcal{B}(n,n)} \langle C, P\rangle$$

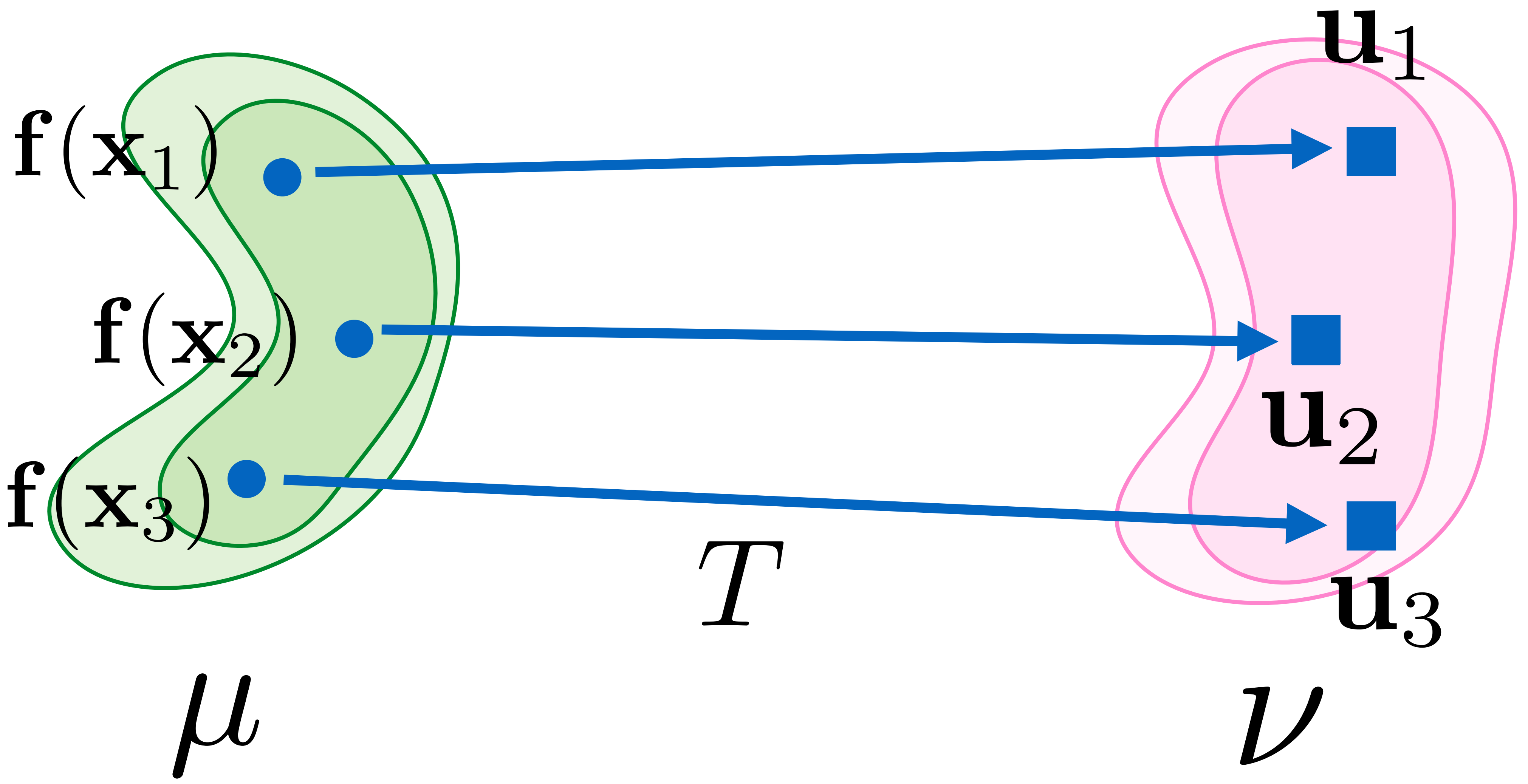


$$C_{ij} = \|\mathbf{f}(\mathbf{x}_i) - \mathbf{u}_j\|^2$$

Brenier isotonic regression

$$\min_{\mathbf{u}_1,\ldots,\mathbf{u}_n} \sum_{i\in[n]} \left\| \mathbf{y}_i - \overbrace{n \sum_{j\in[n]} P^*_{ij}\mathbf{u}_j}^{\text{Barycenteric map } T} \right\|^2$$

$$\text{subject to } P \in \mathop{\arg\min}_{P\in\mathcal{B}(n,n)} \langle C, P\rangle$$



$\mathbf{f}(\mathbf{x}_1)$
$\mathbf{f}(\mathbf{x}_2)$
$\mathbf{f}(\mathbf{x}_3)$
$\mathbf{u}_1$
$\mathbf{u}_2$
$\mathbf{u}_3$
T
μ
ν

$$C_{ij} = \|\mathbf{f}(\mathbf{x}_i) - \mathbf{u}_j\|^2$$

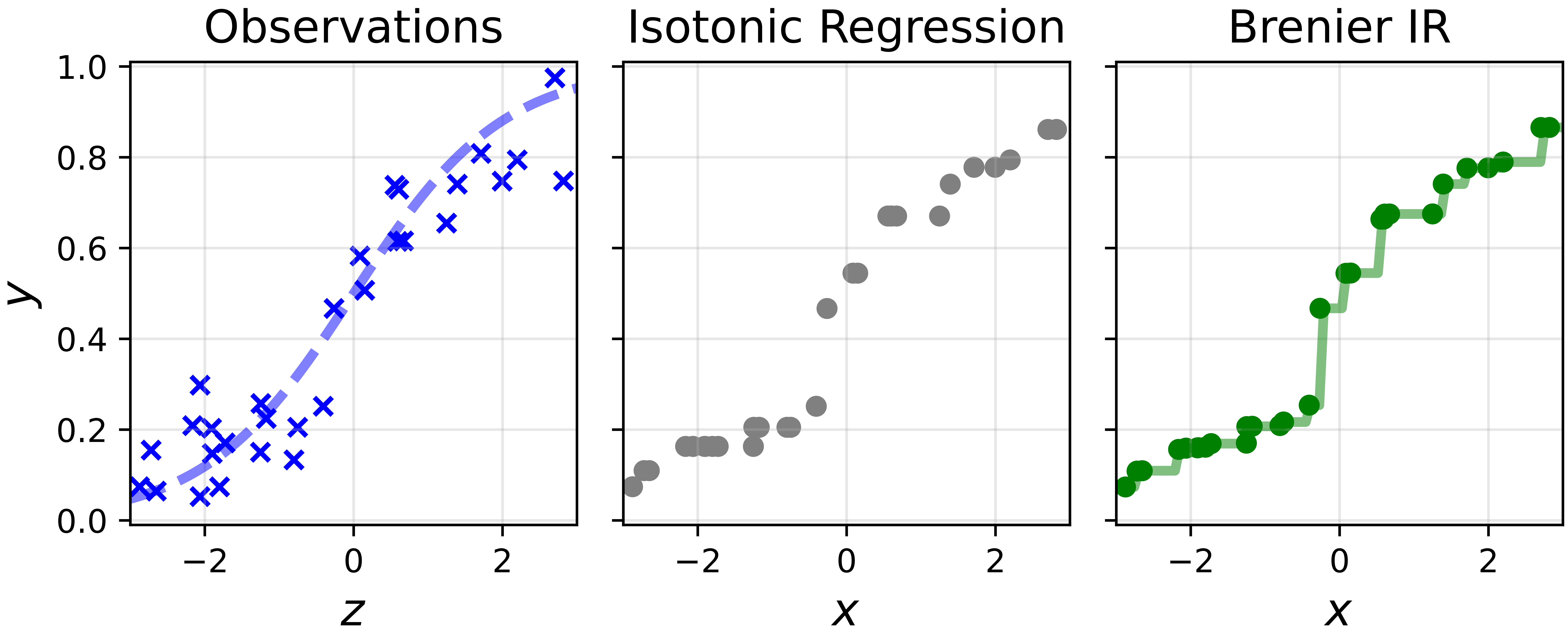
Brenier isotonic regression

$$\boxed{\begin{aligned}&\min_{\mathbf{u}_1,\ldots,\mathbf{u}_n} \boxed{\|Y - nPU\|^2}\\ &\text{subject to } \boxed{P \in \operatorname*{arg\,min}_{P\in\mathcal{B}(n,n)} \langle C, P\rangle}\end{aligned}}$$

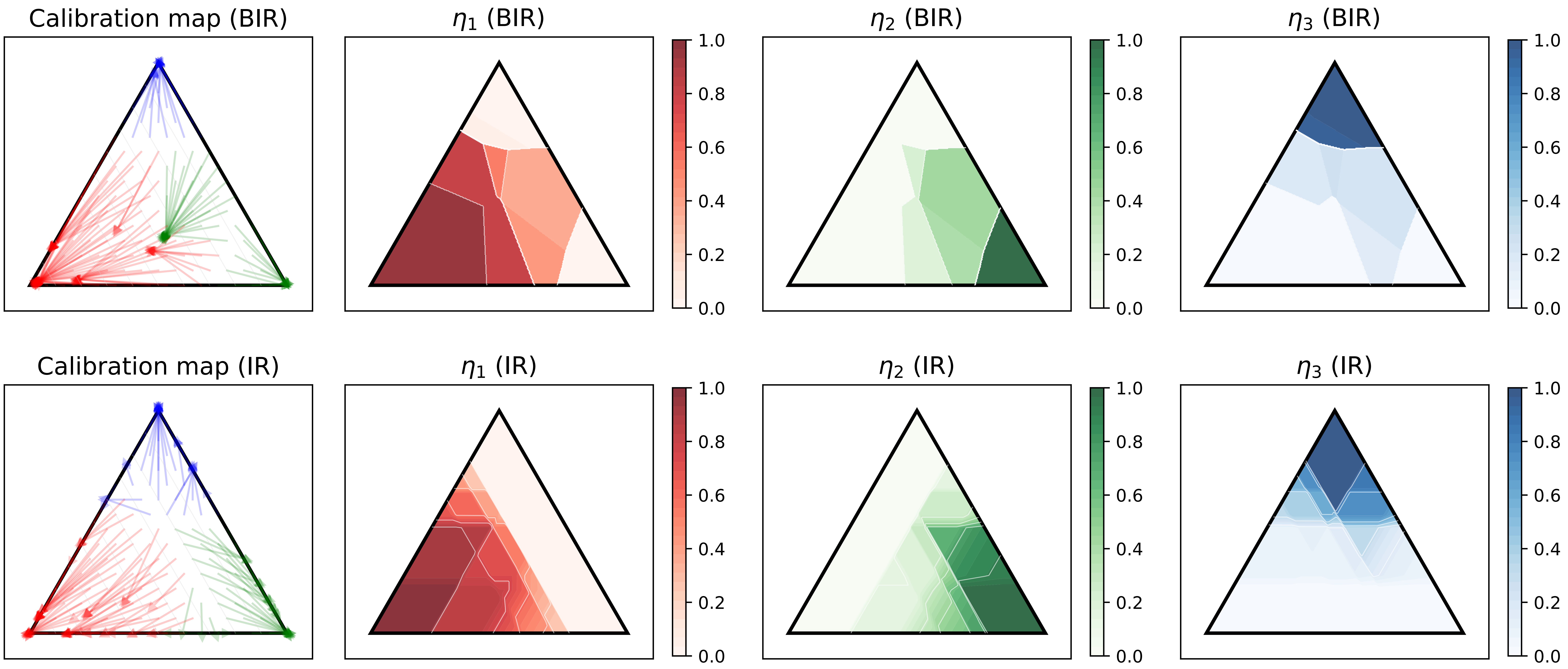
```python
import numpy as np
import ot
from ot.utils.unif
from scipy.optimize import minimize

def obj(u_):
    u = u_.reshape(k, d)
    cost = ot.dist(z, u)
    P = ot.emd(unif(n), unif(k), cost)
    return np.sum((y - n*P@u)**2)/n

res = minimize(obj, [..] method='SLSQP')
u_opt = res.x.reshape(k, d)
```

BIR recovers standard isotonic regression

(of course)

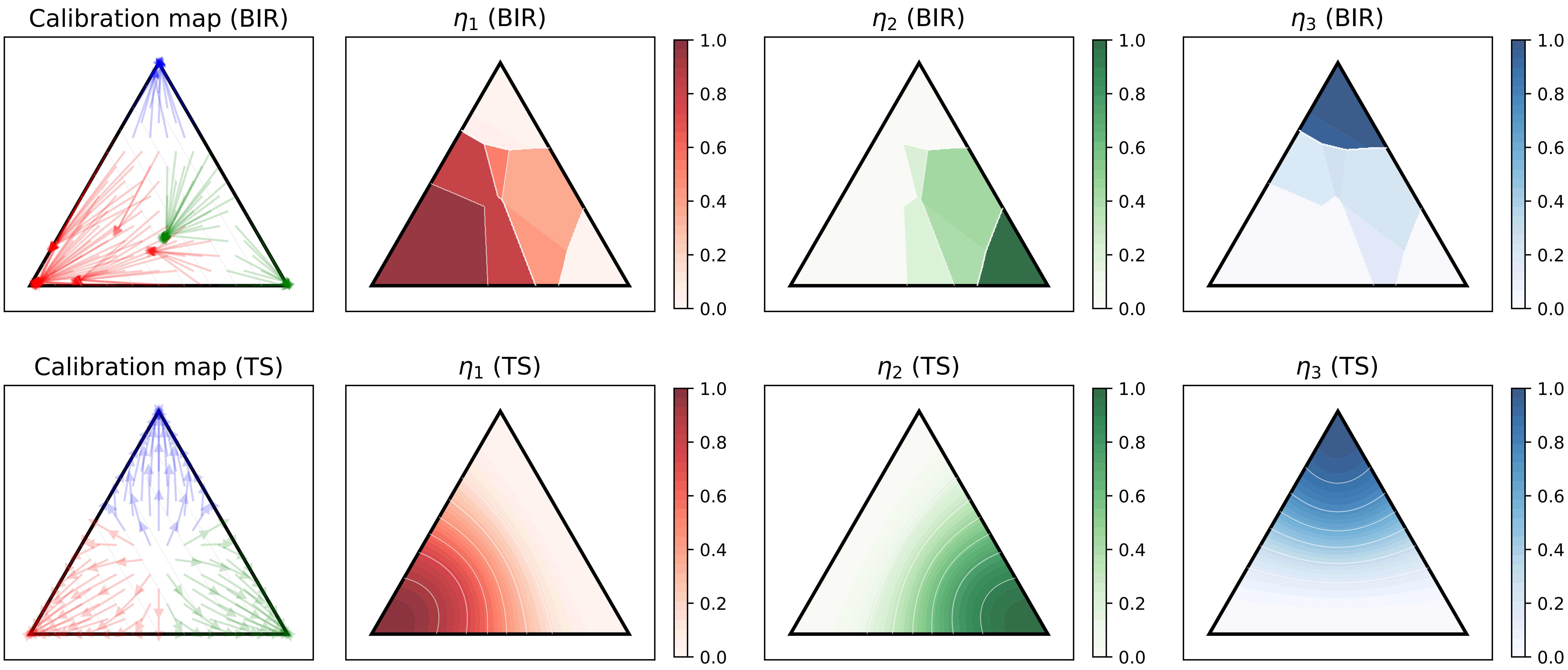


Illustrative examples



Dataset: balance-scale (K=3) / base model: MLP / baseline: one-vs-rest isotonic regression

Illustrative examples



Dataset: balance-scale (K=3) / base model: MLP / baseline: temperature scaling

Benchmark result

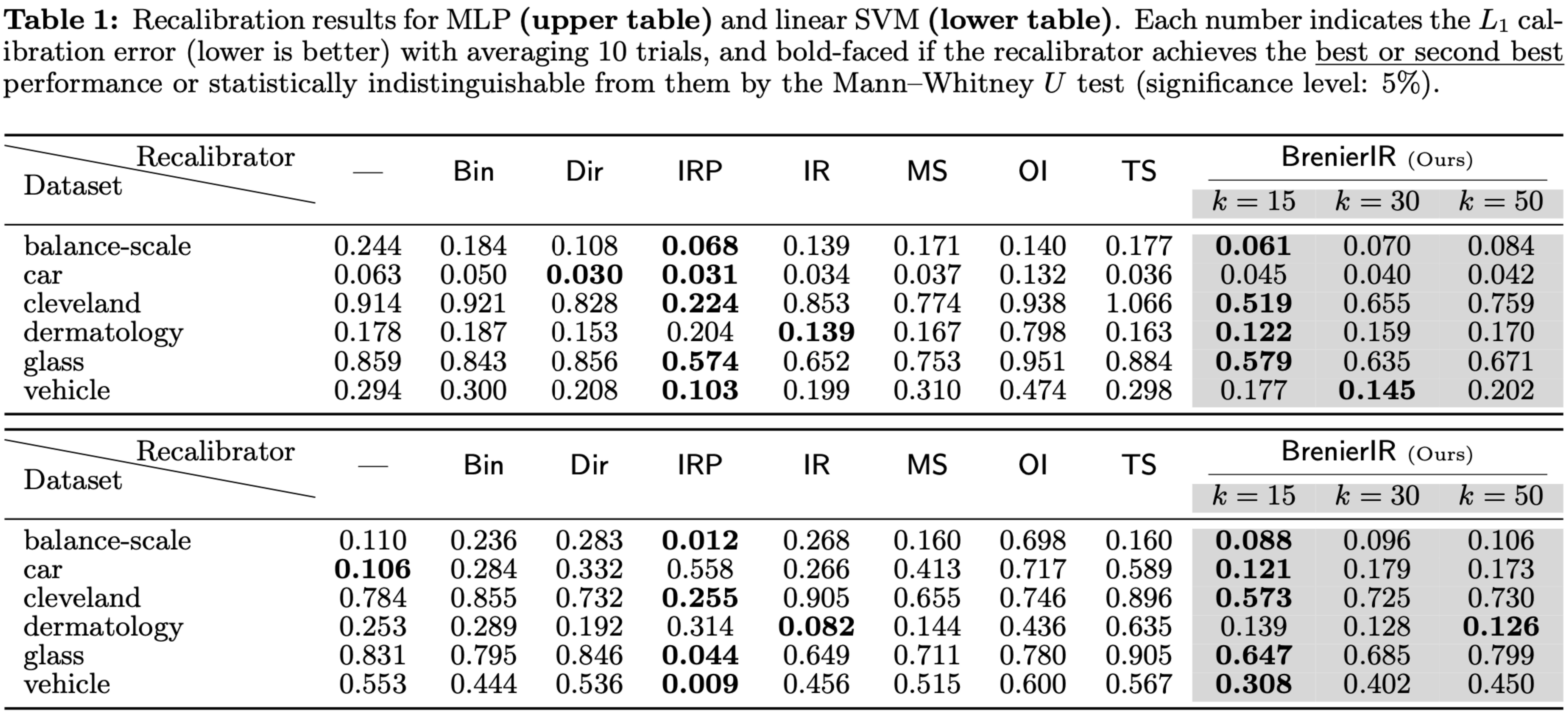


Table 1: Recalibration results for MLP **(upper table)** and linear SVM **(lower table)**. Each number indicates the L_1 calibration error (lower is better) with averaging 10 trials, and bold-faced if the recalibrator achieves the best or second best performance or statistically indistinguishable from them by the Mann–Whitney U test (significance level: 5%).

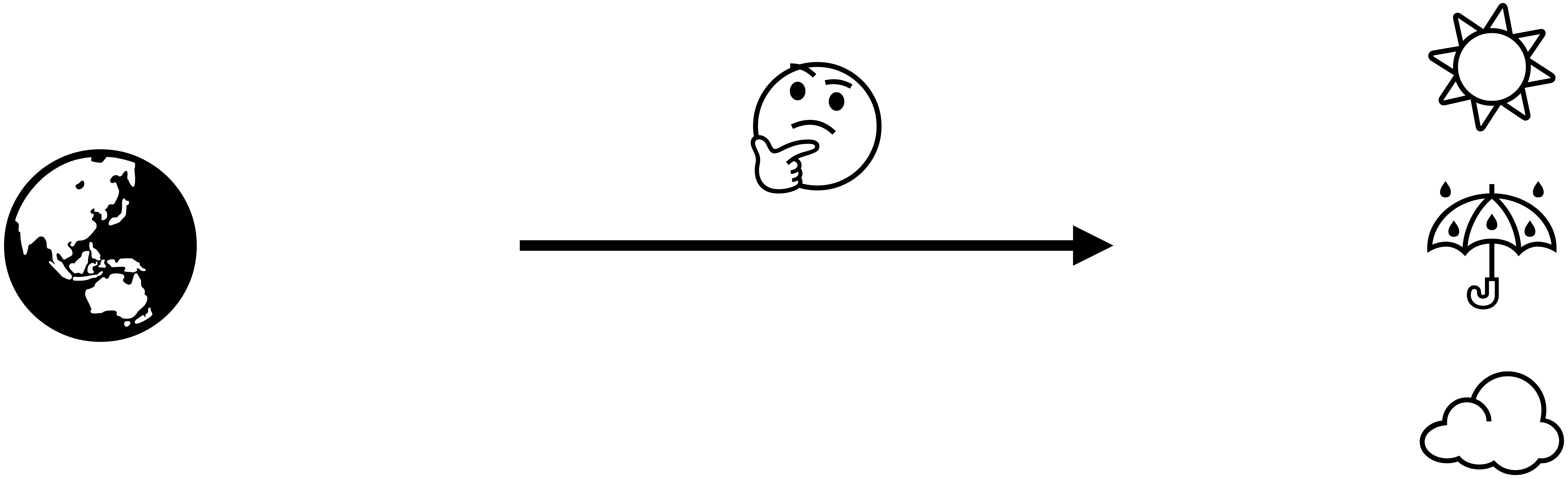
Recalibrator / Dataset	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours)		
									$k = 15$	$k = 30$	$k = 50$
balance-scale	0.244	0.184	0.108	**0.068**	0.139	0.171	0.140	0.177	**0.061**	0.070	0.084
car	0.063	0.050	**0.030**	**0.031**	0.034	0.037	0.132	0.036	0.045	0.040	0.042
cleveland	0.914	0.921	0.828	**0.224**	0.853	0.774	0.938	1.066	**0.519**	0.655	0.759
dermatology	0.178	0.187	0.153	0.204	**0.139**	0.167	0.798	0.163	**0.122**	0.159	0.170
glass	0.859	0.843	0.856	**0.574**	0.652	0.753	0.951	0.884	**0.579**	0.635	0.671
vehicle	0.294	0.300	0.208	**0.103**	0.199	0.310	0.474	0.298	0.177	**0.145**	0.202

Recalibrator / Dataset	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours)		
									$k = 15$	$k = 30$	$k = 50$
balance-scale	0.110	0.236	0.283	**0.012**	0.268	0.160	0.698	0.160	**0.088**	0.096	0.106
car	**0.106**	0.284	0.332	0.558	0.266	0.413	0.717	0.589	**0.121**	0.179	0.173
cleveland	0.784	0.855	0.732	**0.255**	0.905	0.655	0.746	0.896	**0.573**	0.725	0.730
dermatology	0.253	0.289	0.192	0.314	**0.082**	0.144	0.436	0.635	0.139	0.128	**0.126**
glass	0.831	0.795	0.846	**0.044**	0.649	0.711	0.780	0.905	**0.647**	0.685	0.799
vehicle	0.553	0.444	0.536	**0.009**	0.456	0.515	0.600	0.567	**0.308**	0.402	0.450

One step further

Quantify closeness of class probability estimates

Multiclass classification



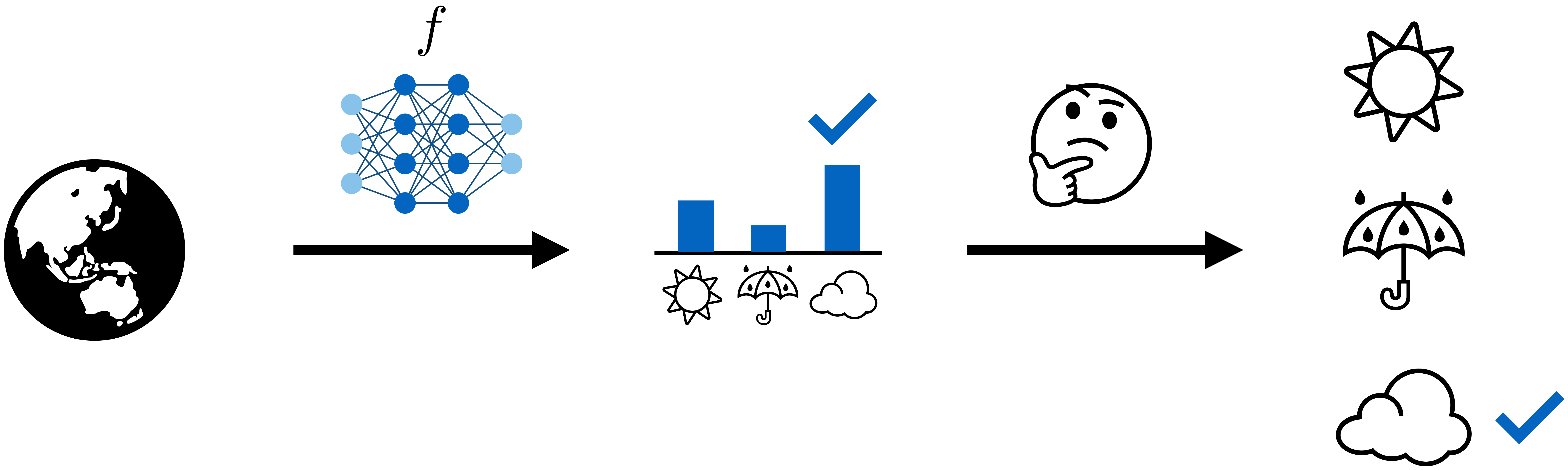
x

feature

$\widehat{y}$

class

Multiclass classification



$\boldsymbol{x}$

feature

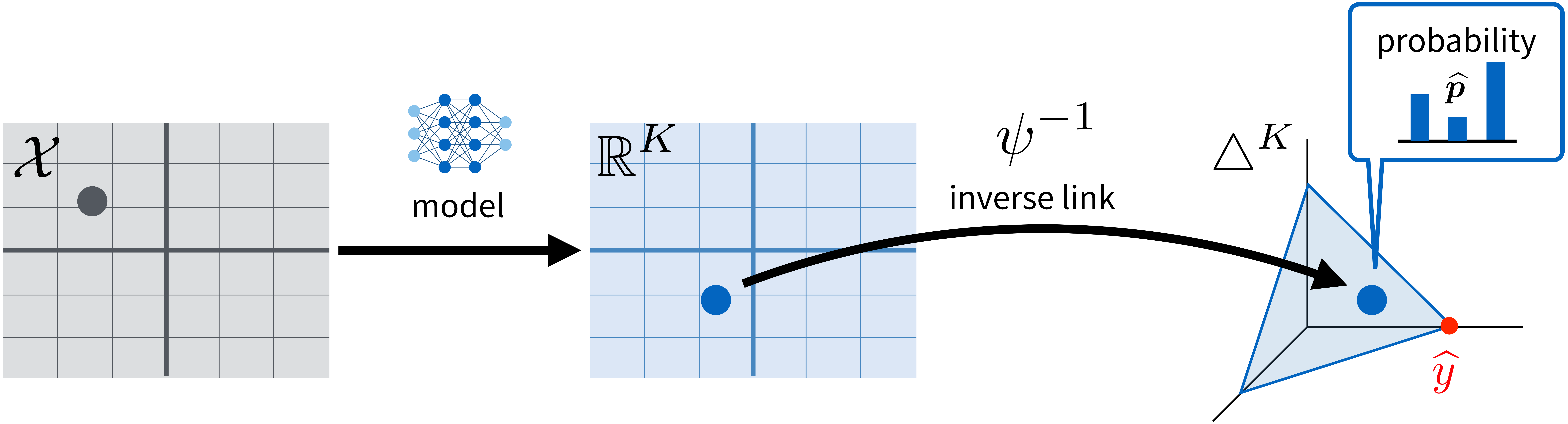
$\boldsymbol{r} = f(\boldsymbol{x})$

report

$\widehat{y}$

class

Multiclass classification



$\mathcal{X}$

model

$\mathbb{R}^K$

ψ^{-1}

inverse link

$\triangle^K$

probability

$\widehat{\boldsymbol{p}}$

$\widehat{y}$

$$\boldsymbol{x}$$

feature

$$\boldsymbol{r} = f(\boldsymbol{x})$$

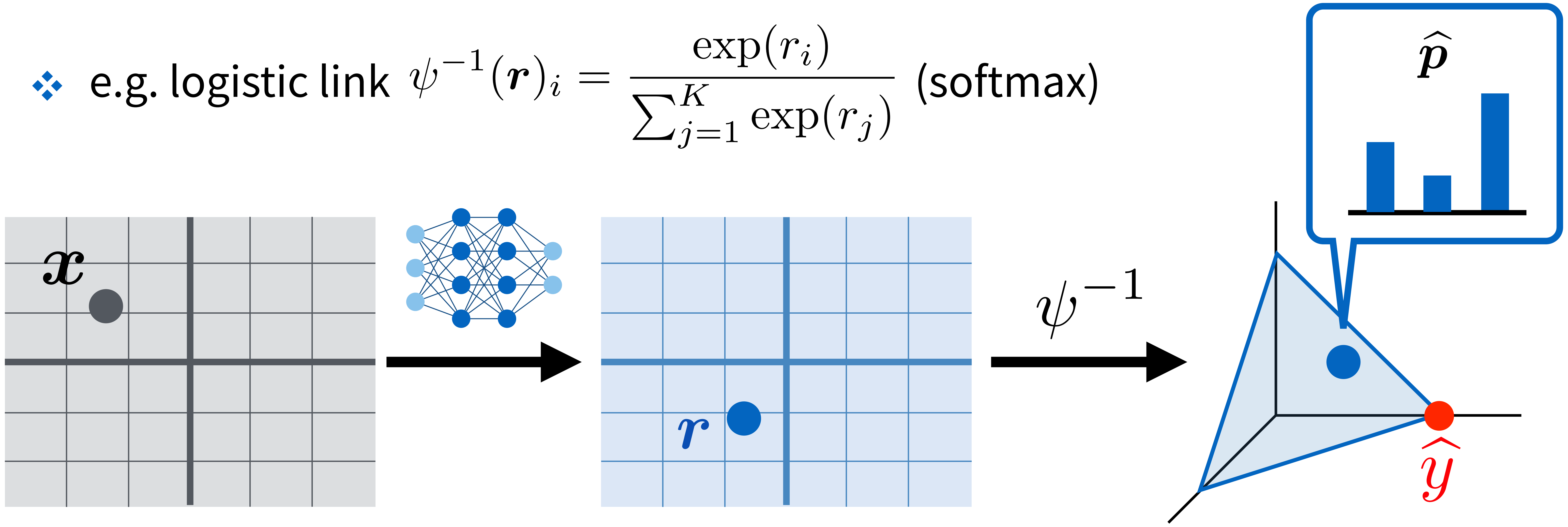
report

$$\widehat{y}$$

class

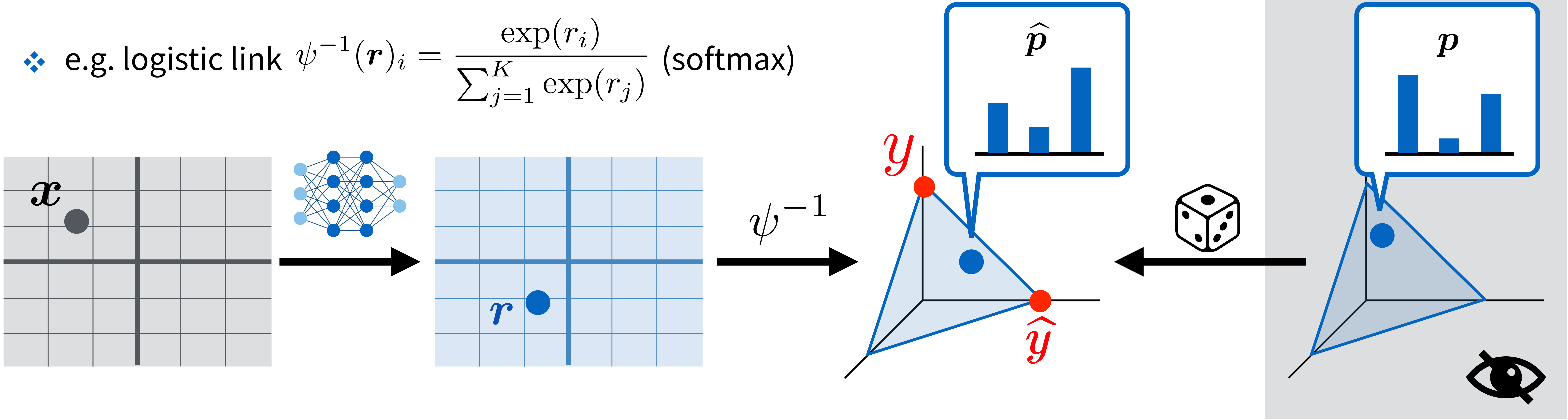
Class probability estimation

- Define report space $\mathbb{R}^K$ for K-class classification
- Inverse link function ψ^{-1} maps report $\boldsymbol{r} \in \mathbb{R}^K$ to prediction $\widehat{\boldsymbol{p}} \in \triangle^K$
  - e.g. logistic link $\psi^{-1}(\boldsymbol{r})_i = \dfrac{\exp(r_i)}{\sum_{j=1}^{K} \exp(r_j)}$ (softmax)



Class probability estimation

- Define report space $\mathbb{R}^K$ for K-class classification
- Inverse link function ψ^{-1} maps report $\boldsymbol{r} \in \mathbb{R}^K$ to prediction $\widehat{\boldsymbol{p}} \in \triangle^K$
  - e.g. logistic link $\psi^{-1}(\boldsymbol{r})_i = \dfrac{\exp(r_i)}{\sum_{j=1}^K \exp(r_j)}$ (softmax)



How to measure $\mathrm{dist}(\boldsymbol{p}, \widehat{\boldsymbol{p}})$?

Proper loss

loss functions for class probability estimation

Loss function (scoring rule)

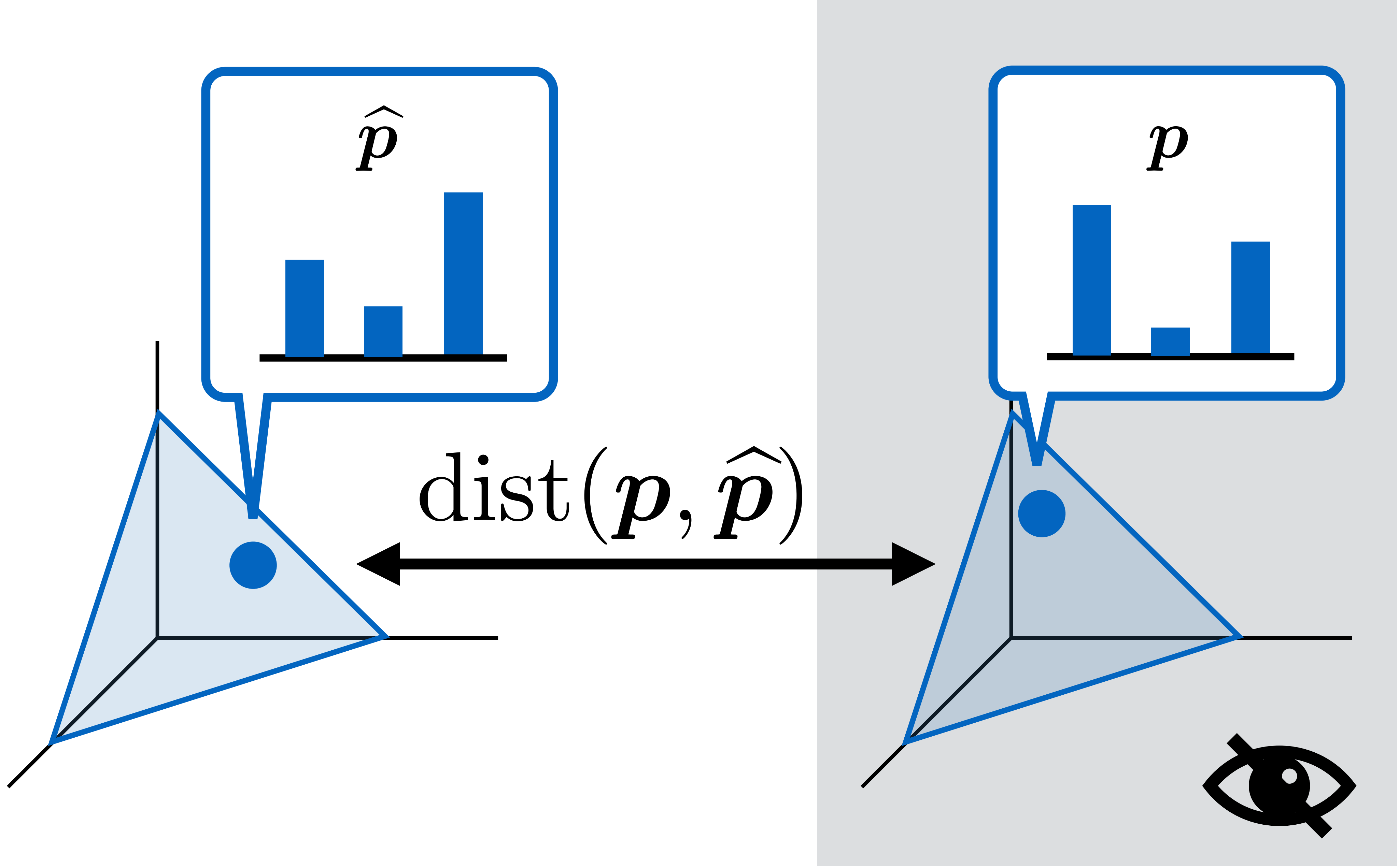
[Buja+ 2005]

- Define loss function $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$

Definition Conditional risk

$$L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \sum_{y=1}^{K} p_y \ell_y(\widehat{\boldsymbol{p}}) = \langle \boldsymbol{p}, \boldsymbol{\ell}(\widehat{\boldsymbol{p}}) \rangle$$

- expected loss under the true probability $\boldsymbol{p}$


$\widehat{\boldsymbol{p}}$
$\boldsymbol{p}$
$\mathrm{dist}(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

Andreas Buja, Werner Stuetzle, and Yi Shen.
Loss functions for binary class probability estimation and classification: Structure and applications. Technical Report, 2005.

Loss function (scoring rule)

[Buja+ 2005]

- Define loss function $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$

Definition Conditional risk

$$L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \langle \boldsymbol{p}, \boldsymbol{\ell}(\widehat{\boldsymbol{p}}) \rangle$$

- Empirical risk minimization

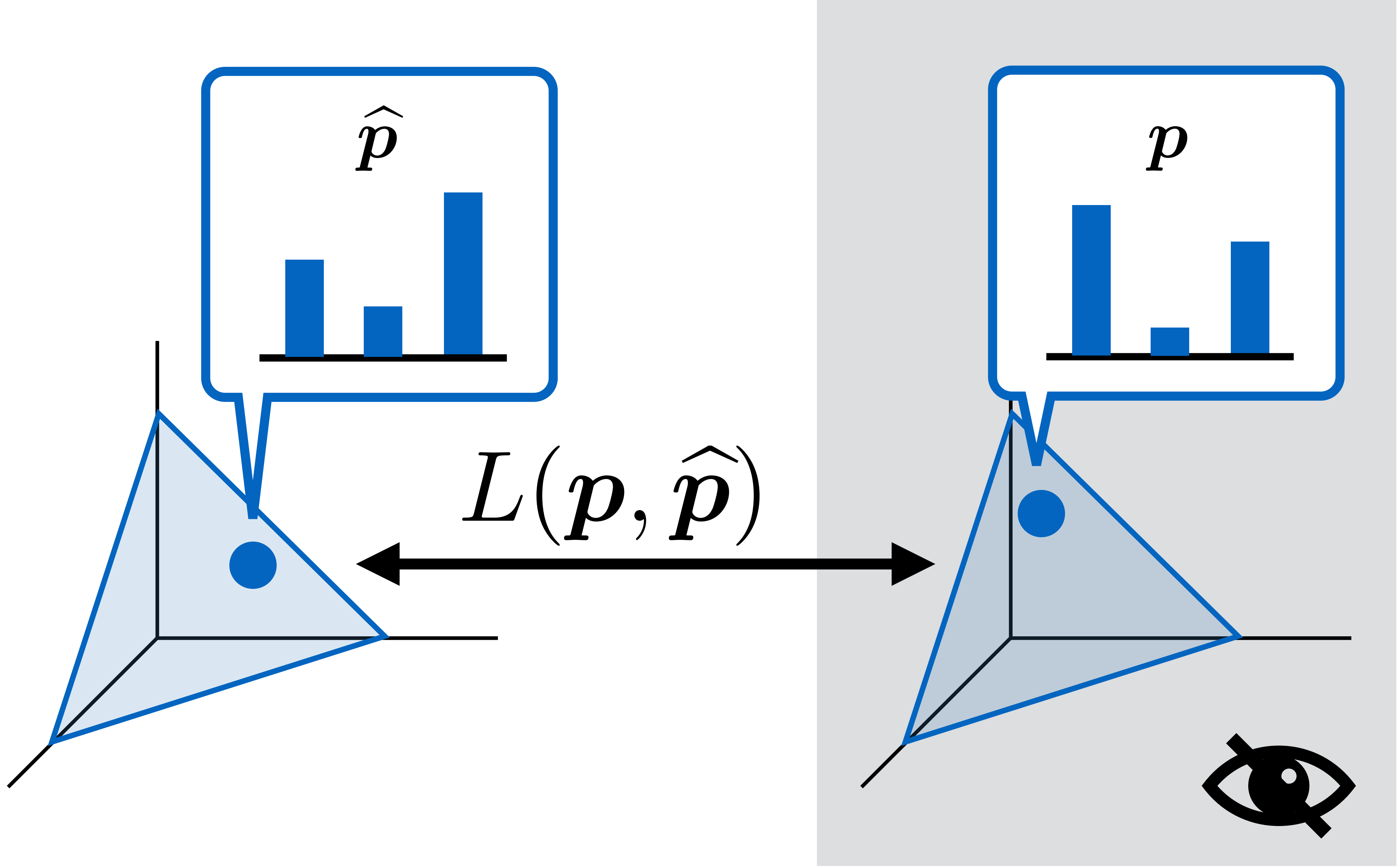
$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell_{y_n}(\psi^{-1}(f(\boldsymbol{x}_n)))$$

⬆target: label

uniform convergence

$$R(f) = \mathbb{E}_X \left[L\big(\mathbb{P}(Y|X), \psi^{-1}(f(X))\big) \right]$$

⬆target: class prob


$\widehat{\boldsymbol{p}}$
$\boldsymbol{p}$
$L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

Andreas Buja, Werner Stuetzle, and Yi Shen.
Loss functions for binary class probability estimation and classification: Structure and applications. Technical Report, 2005.

Loss function (scoring rule)

[Buja+ 2005]

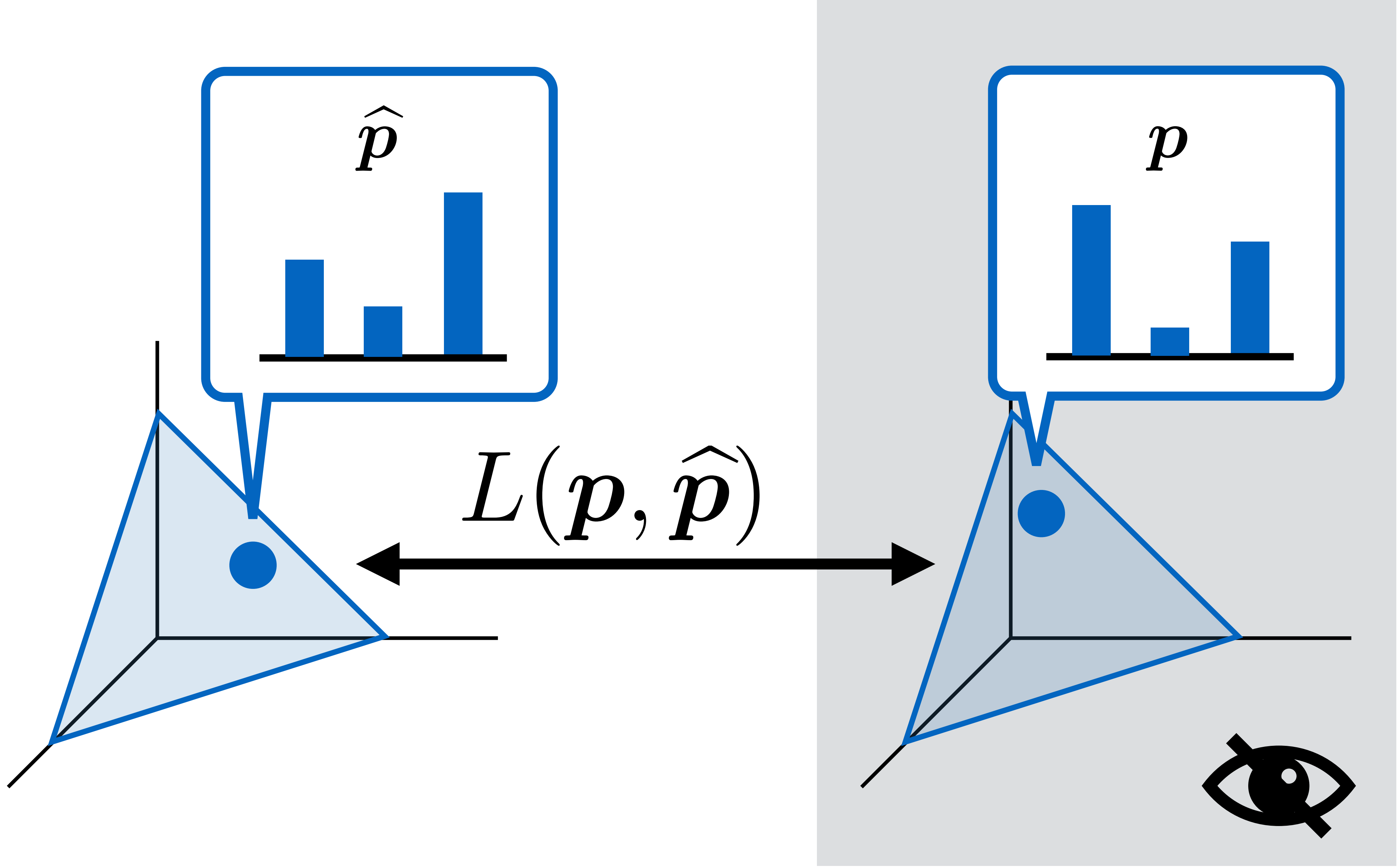
- Define loss function $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$

Definition Conditional risk

$$L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \langle \boldsymbol{p}, \boldsymbol{\ell}(\widehat{\boldsymbol{p}}) \rangle$$

Definition Bayes risk $\underline{L}(\boldsymbol{p}) = \inf_{\widehat{\boldsymbol{p}} \in \triangle^K} L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

- best possible loss under the true probability $\boldsymbol{p}$


$\widehat{\boldsymbol{p}}$
$\boldsymbol{p}$
$L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

Andreas Buja, Werner Stuetzle, and Yi Shen.
Loss functions for binary class probability estimation and classification: Structure and applications. Technical Report, 2005.

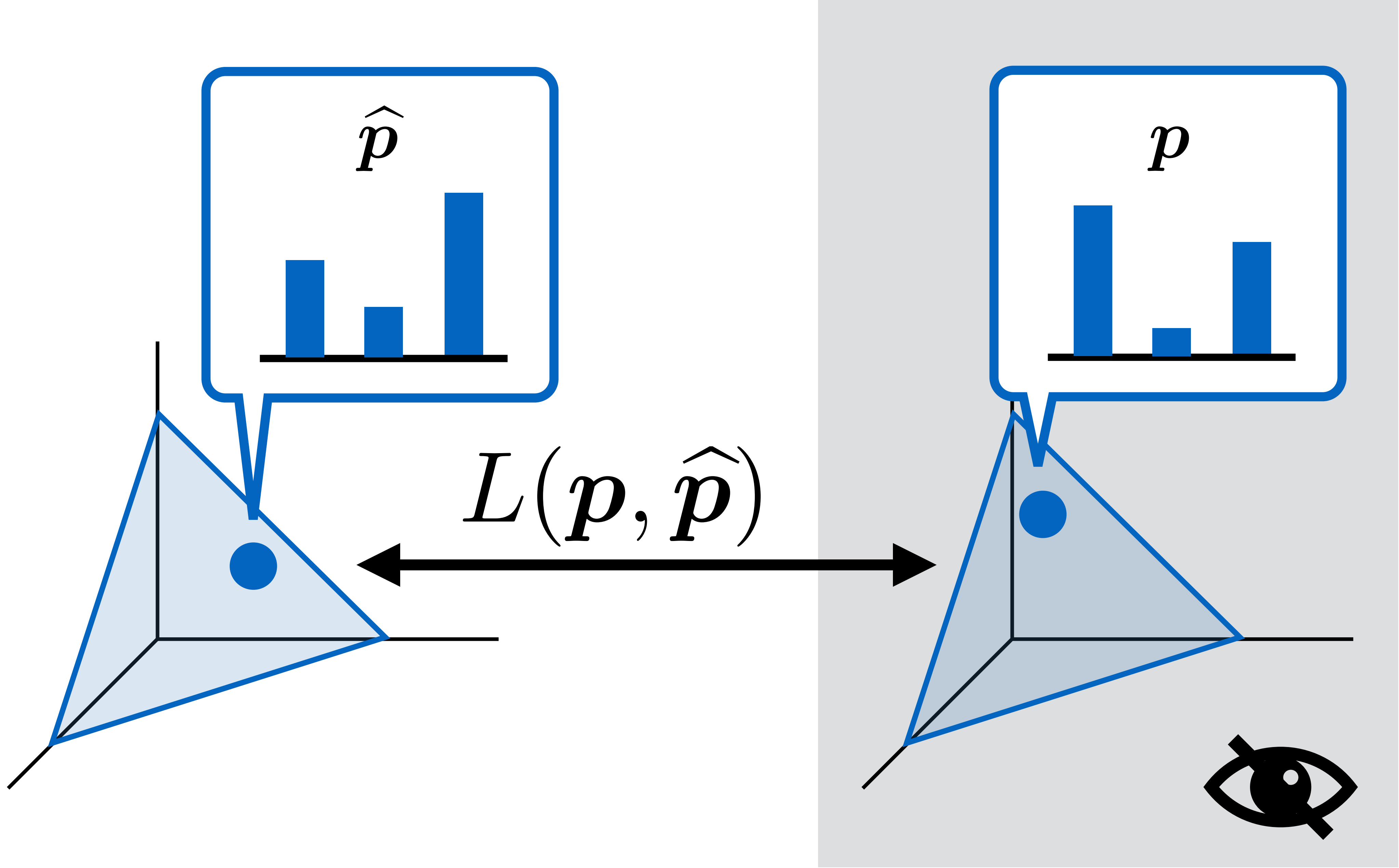
Loss function (scoring rule)

[Buja+ 2005]

Definition $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \langle \boldsymbol{p}, \boldsymbol{\ell}(\widehat{\boldsymbol{p}}) \rangle$ $\quad\quad$ $\underline{L}(\boldsymbol{p}) = \inf_{\widehat{\boldsymbol{p}} \in \triangle^K} L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

Q What is an admissible loss function?

A It is minimized at true class probability


$\widehat{\boldsymbol{p}}$
$\boldsymbol{p}$
$L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

Andreas Buja, Werner Stuetzle, and Yi Shen.
Loss functions for binary class probability estimation and classification: Structure and applications. Technical Report, 2005.

Proper loss (a.k.a. proper scoring rule)

[Buja+ 2005]

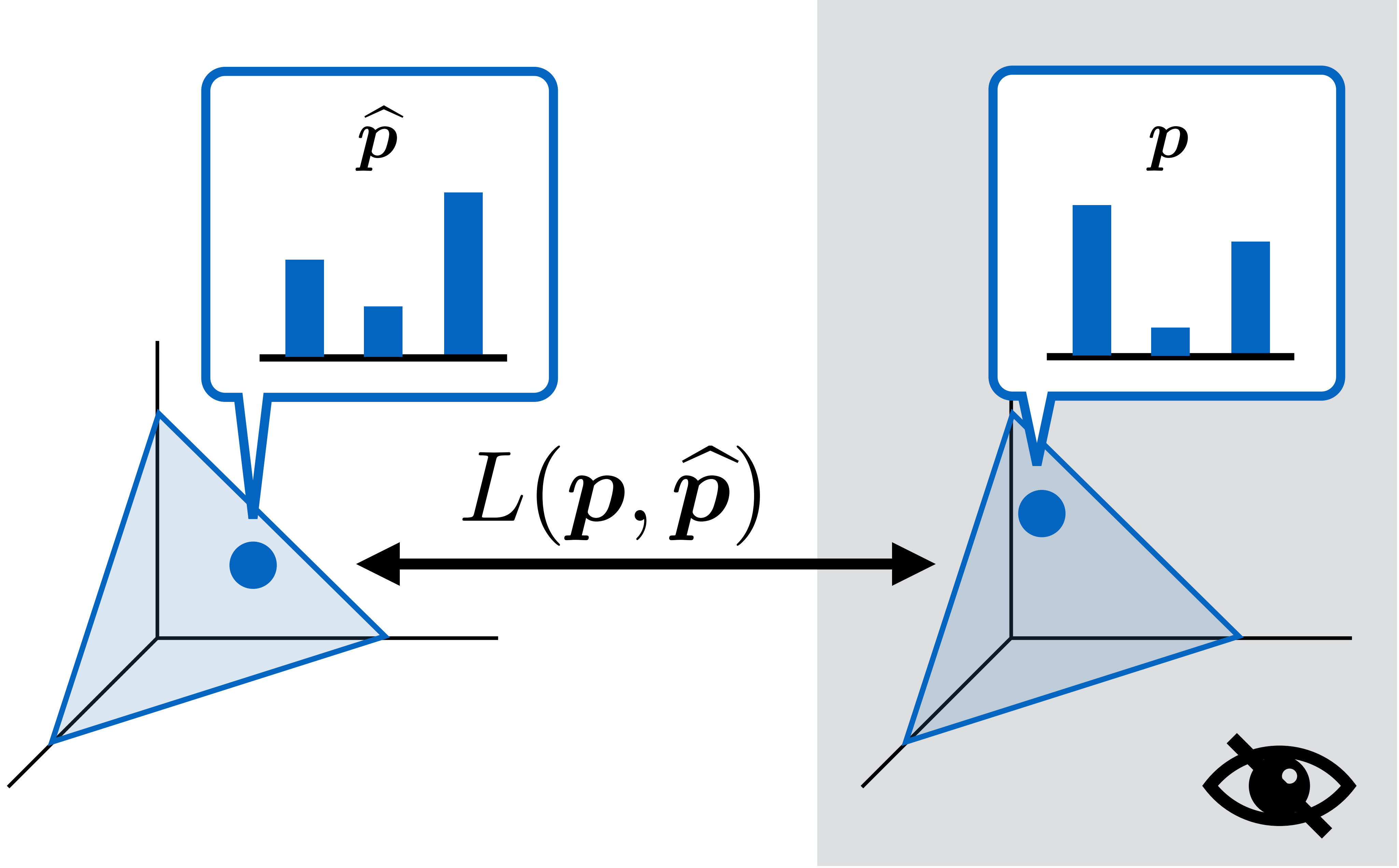
Definition $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \langle \boldsymbol{p}, \boldsymbol{\ell}(\widehat{\boldsymbol{p}}) \rangle$ $\quad$ $\underline{L}(\boldsymbol{p}) = \inf_{\widehat{\boldsymbol{p}} \in \triangle^K} L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

Q What is an admissible loss function?

A It is minimized at true class probability

Definition Loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is strictly proper if

$$L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) > L(\boldsymbol{p}, \boldsymbol{p}) = \underline{L}(\boldsymbol{p}) \text{ for all } \boldsymbol{p} \neq \widehat{\boldsymbol{p}}.$$


$\widehat{\boldsymbol{p}}$
$\boldsymbol{p}$
$L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

Andreas Buja, Werner Stuetzle, and Yi Shen.
Loss functions for binary class probability estimation and classification: Structure and applications. Technical Report, 2005.

Proper loss is Bregman divergence

[Buja+ 2005]

Definition $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \langle \boldsymbol{p}, \boldsymbol{\ell}(\widehat{\boldsymbol{p}}) \rangle$ $\quad \underline{L}(\boldsymbol{p}) = \inf_{\widehat{\boldsymbol{p}} \in \triangle^K} L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

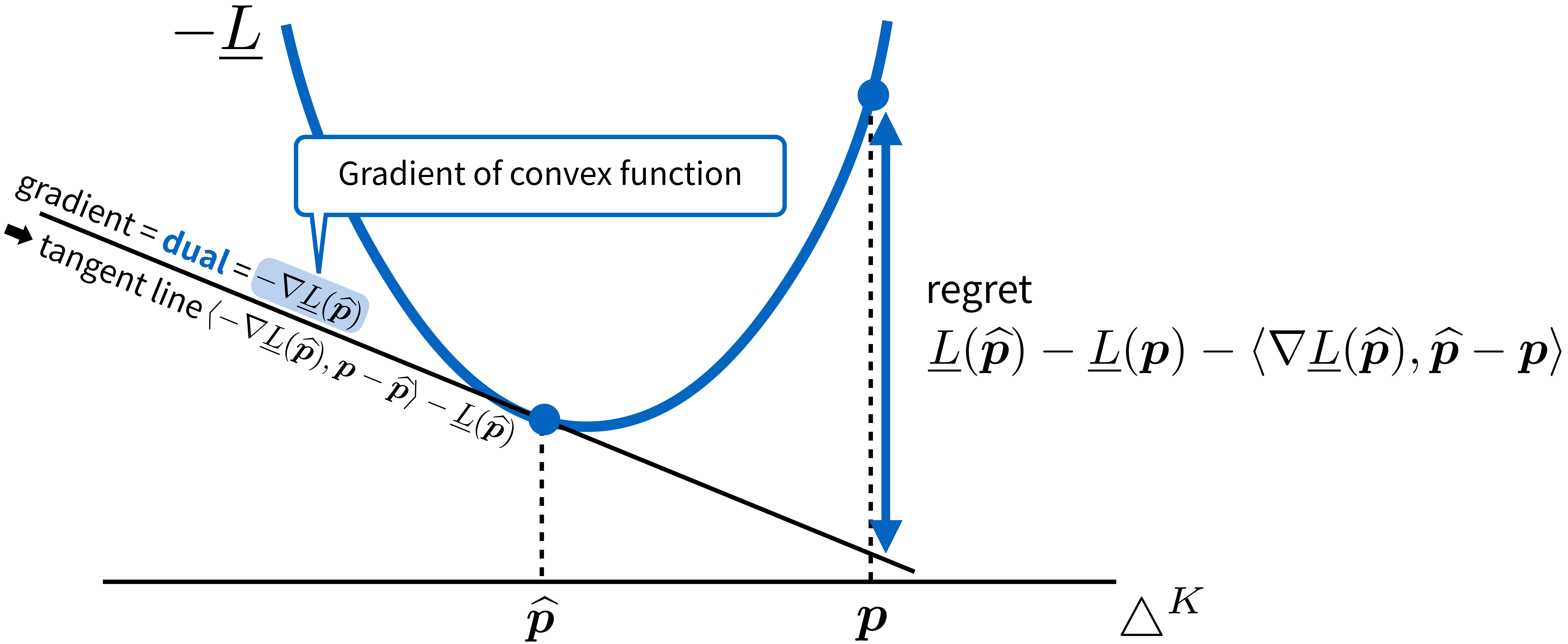
Definition Loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is strictly proper if $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) > L(\boldsymbol{p}, \boldsymbol{p}) = \underline{L}(\boldsymbol{p})$ for all $\boldsymbol{p} \neq \widehat{\boldsymbol{p}}$.

Theorem A regular loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is strictly proper if and only if $-\underline{L}$ is strictly convex, and for all $\boldsymbol{p}, \widehat{\boldsymbol{p}} \in \triangle^K$, the regret satisfies

$$\underbrace{L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) - \underline{L}(\boldsymbol{p})}_{\text{regret}} = \underbrace{\underline{L}(\widehat{\boldsymbol{p}}) - \underline{L}(\boldsymbol{p}) - \langle \nabla \underline{L}(\widehat{\boldsymbol{p}}), \widehat{\boldsymbol{p}} - \boldsymbol{p} \rangle.}_{\text{Bregman divergence generated by } -\underline{L}}$$

Re-discovered many times:
McCarthy (1956), Savage (1971), Buja et al. (2005), Gneiting and Raftery (2007), Reid and Williamson (2010), etc.

Andreas Buja, Werner Stuetzle, and Yi Shen.
Loss functions for binary class probability estimation and classification: Structure and applications. Technical Report, 2005.



$-\underline{L}$

Gradient of convex function

gradient = **dual** = $-\nabla\underline{L}(\widehat{\boldsymbol{p}})$

➡ tangent line $\langle -\nabla\underline{L}(\widehat{\boldsymbol{p}}), \boldsymbol{p} - \widehat{\boldsymbol{p}}\rangle - \underline{L}(\widehat{\boldsymbol{p}})$

regret

$\underline{L}(\widehat{\boldsymbol{p}}) - \underline{L}(\boldsymbol{p}) - \langle \nabla\underline{L}(\widehat{\boldsymbol{p}}), \widehat{\boldsymbol{p}} - \boldsymbol{p}\rangle$

$\widehat{\boldsymbol{p}}$

$\boldsymbol{p}$

$\triangle^K$

Proper loss as a Bregman divergence in **primal** space

Example

Definition $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \langle \boldsymbol{p}, \boldsymbol{\ell}(\widehat{\boldsymbol{p}}) \rangle$ $\qquad$ $\underline{L}(\boldsymbol{p}) = \inf_{\widehat{\boldsymbol{p}} \in \triangle^K} L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

Definition Loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is strictly proper if $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) > L(\boldsymbol{p}, \boldsymbol{p}) = \underline{L}(\boldsymbol{p})$ for all $\boldsymbol{p} \neq \widehat{\boldsymbol{p}}$.

- Log loss $\ell_y(\widehat{\boldsymbol{p}}) = -\ln \widehat{p}_y$
  - Conditional risk $\quad L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = -\langle \boldsymbol{p}, \ln \widehat{\boldsymbol{p}} \rangle$ $\quad$ (cross entropy)
  - Bayes risk $\quad \underline{L}(\boldsymbol{p}) = -\langle \boldsymbol{p}, \ln \boldsymbol{p} \rangle$ $\quad$ (Shannon entropy)
  - Regret $\quad L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) - \underline{L}(\boldsymbol{p}) = \left\langle \boldsymbol{p}, \ln \frac{\boldsymbol{p}}{\widehat{\boldsymbol{p}}} \right\rangle$ $\quad$ (Kullback-Leibler divergence)

Example

Definition $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \langle \boldsymbol{p}, \boldsymbol{\ell}(\widehat{\boldsymbol{p}}) \rangle$ $\underline{L}(\boldsymbol{p}) = \inf_{\widehat{\boldsymbol{p}} \in \triangle^K} L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$

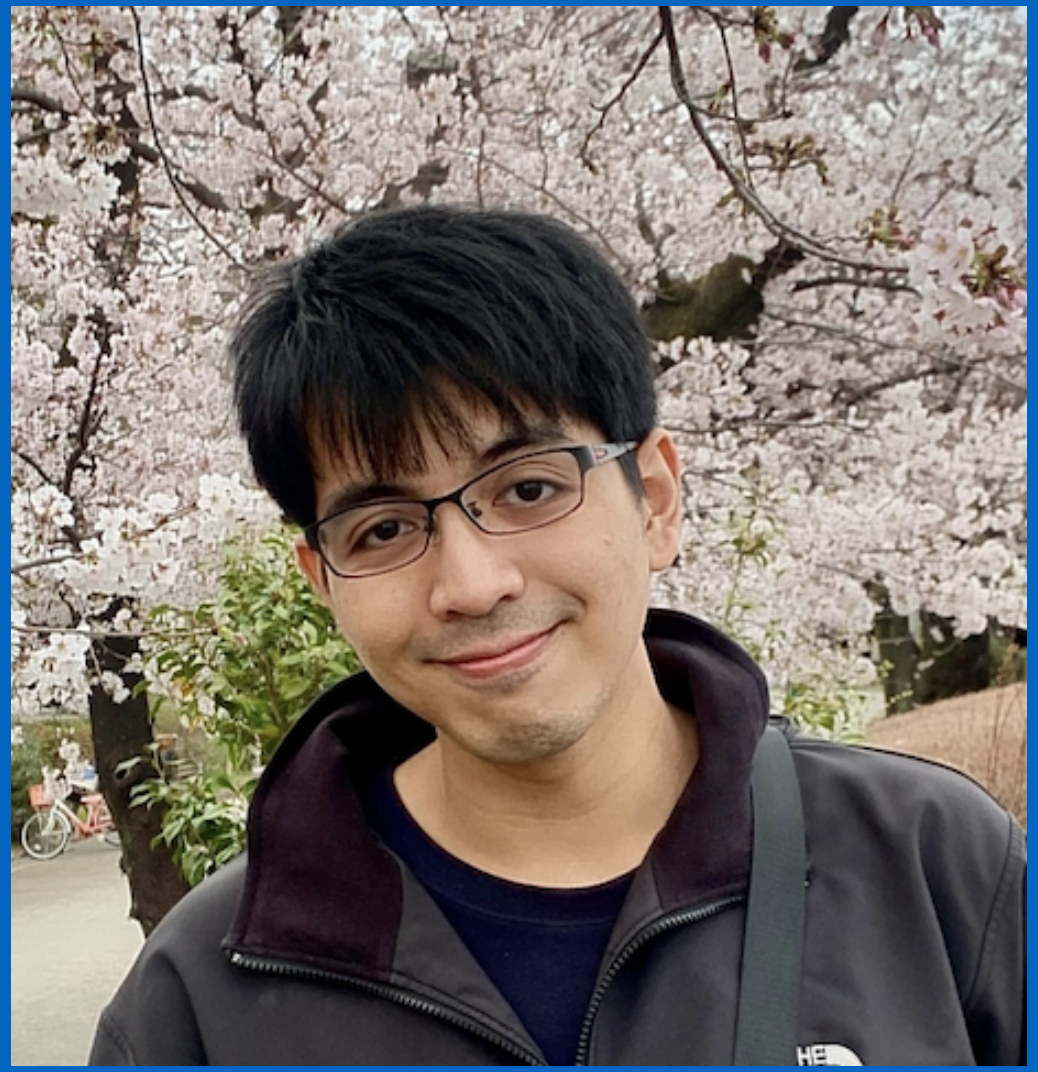
Definition Loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is strictly proper if $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) > L(\boldsymbol{p}, \boldsymbol{p}) = \underline{L}(\boldsymbol{p})$ for all $\boldsymbol{p} \neq \widehat{\boldsymbol{p}}$.

- Brier loss $\ell_y(\widehat{\boldsymbol{p}}) = -\widehat{p}_y + (1 + \|\widehat{\boldsymbol{p}}\|_2^2)/2$
  - Conditional risk $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \dfrac{1 - 2\langle \boldsymbol{p}, \widehat{\boldsymbol{p}} \rangle + \|\widehat{\boldsymbol{p}}\|_2^2}{2}$
  - Bayes risk $\underline{L}(\boldsymbol{p}) = \dfrac{1 - \|\boldsymbol{p}\|_2^2}{2}$ (Gini index)
  - Regret $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) - \underline{L}(\boldsymbol{p}) = \dfrac{1}{2}\|\boldsymbol{p} - \widehat{\boldsymbol{p}}\|_2^2$ (squared L2 distance)

Calm Composite Losses

Being Improper Yet Proper Composite

Joint work with Nontawat Charoenphakdee,
and presented at AISTATS2025



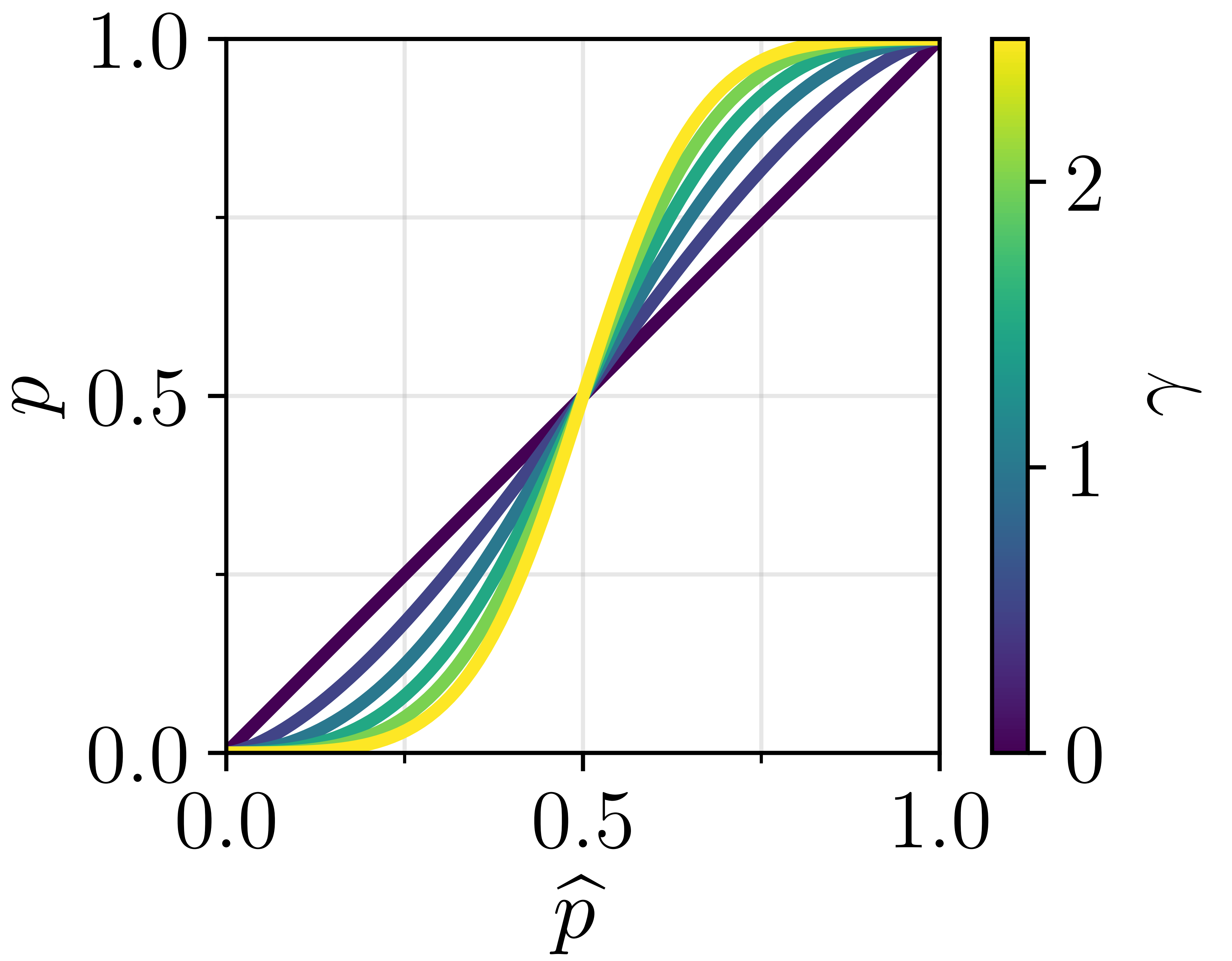
Commonly used losses are not necessarily proper

- **Example.** Focal loss $\ell_y(\widehat{\boldsymbol{p}}) = -(1-\widehat{p}_y)^\gamma \ln \widehat{p}_y$ [Lin+ 2017]
  - motivation: downweight overconfident class probability for imbalanced problems

- 😥 **Focal loss is improper**

Definition Loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is strictly proper if $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) > L(\boldsymbol{p}, \boldsymbol{p}) = \underline{L}(\boldsymbol{p})$ for all $\boldsymbol{p} \neq \widehat{\boldsymbol{p}}$.

  - for binary classification, plot optimal $\widehat{p}$ of conditional risk (➡)
  - diagonal line is expected



Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 2980–2988, 2017.

Commonly used losses are not necessarily proper

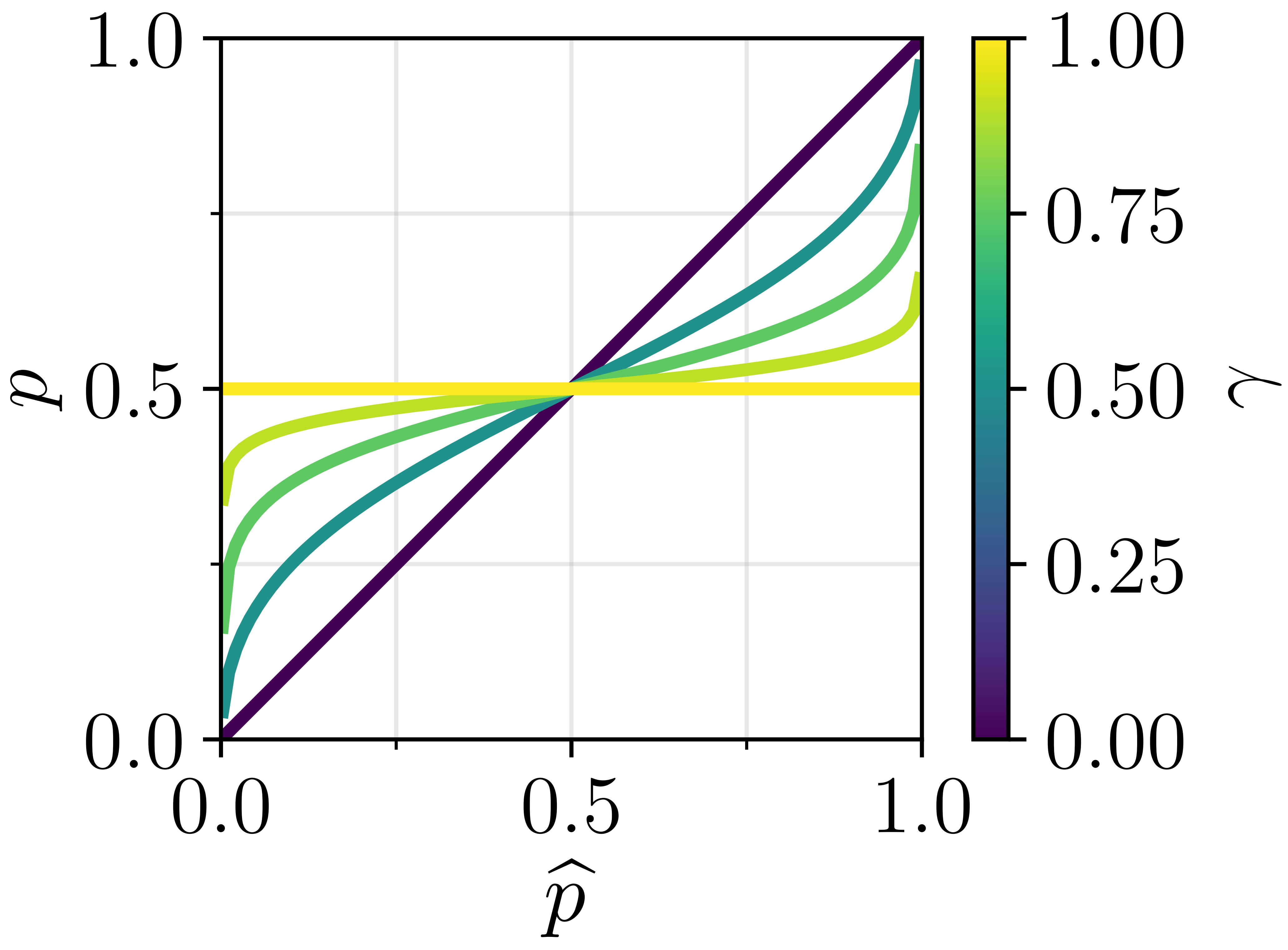
[Zhang & Sabuncu 2018]

- **Example.** Generalized cross-entropy loss $\ell_y(\widehat{\boldsymbol{p}}) = (1 - \widehat{p}_y^{\gamma})/\gamma$
  - motivation: interpolate log loss (not robust) and MAE loss (robust)

- 😥 **Generalized cross-entropy loss is improper**

Definition Loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is strictly proper if $L(\boldsymbol{p}, \widehat{\boldsymbol{p}}) > L(\boldsymbol{p}, \boldsymbol{p}) = \underline{L}(\boldsymbol{p})$ for all $\boldsymbol{p} \neq \widehat{\boldsymbol{p}}$.

  - for binary classification, plot optimal $\widehat{p}$ of conditional risk (➡)
  - diagonal line is expected

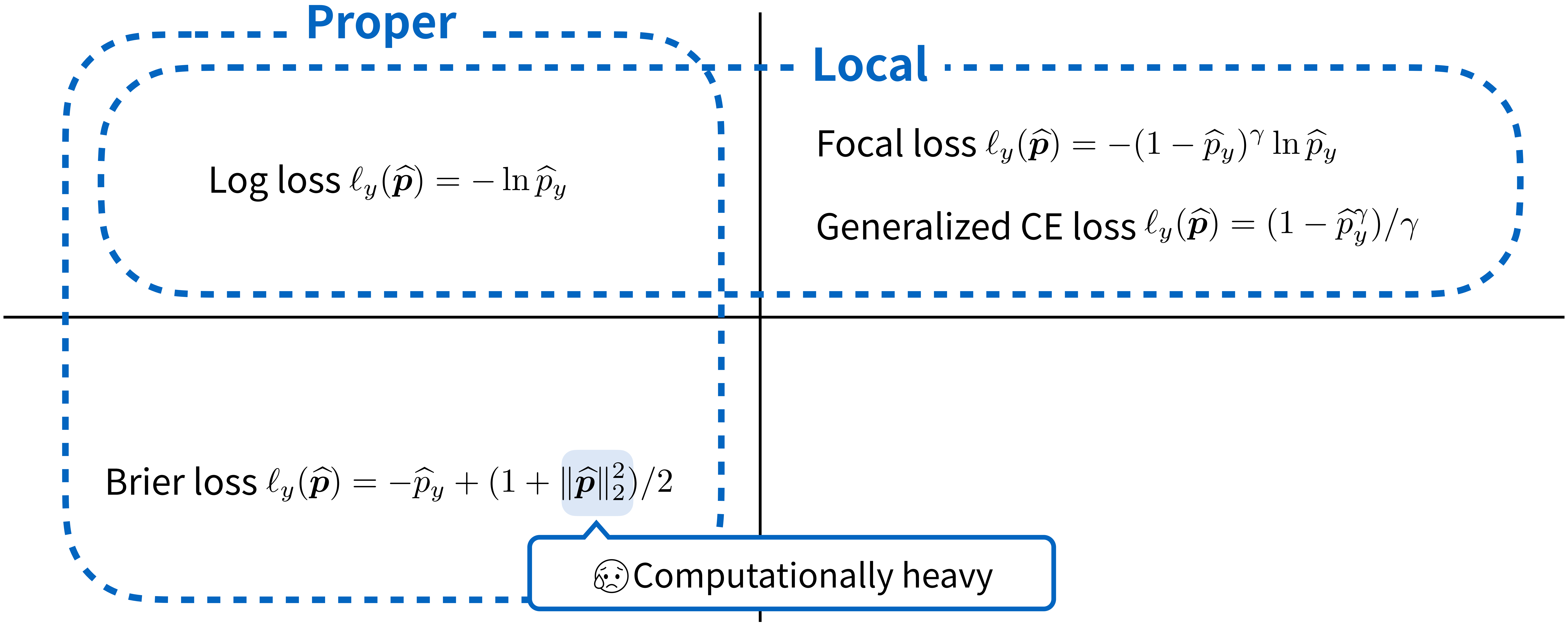


Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in Neural Information Processing Systems, 31:8778–8788, 2018.

Remark: Log loss is the only “local” proper loss

- Proper loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is **local** if $\ell_y(\widehat{\boldsymbol{p}})$ depends on $\widehat{p}_y$ solely

[Parry+ 2012]



Proper

Local

Log loss $\ell_y(\widehat{\boldsymbol{p}}) = -\ln \widehat{p}_y$

Focal loss $\ell_y(\widehat{\boldsymbol{p}}) = -(1 - \widehat{p}_y)^\gamma \ln \widehat{p}_y$

Generalized CE loss $\ell_y(\widehat{\boldsymbol{p}}) = (1 - \widehat{p}_y^\gamma)/\gamma$

Brier loss $\ell_y(\widehat{\boldsymbol{p}}) = -\widehat{p}_y + (1 + \|\widehat{\boldsymbol{p}}\|_2^2)/2$

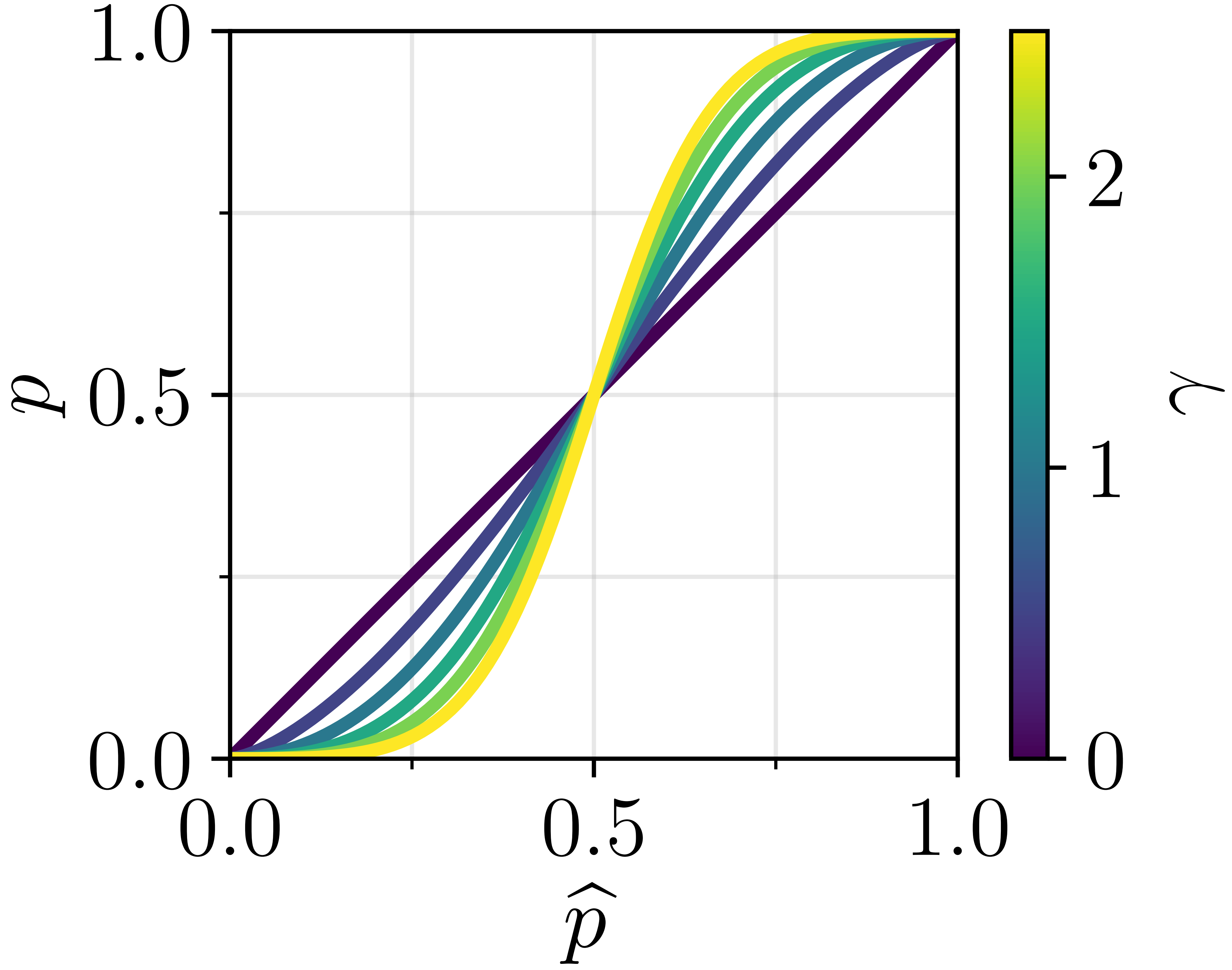
😥Computationally heavy

Mattew Parry, A Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. The Annals of Statistics, 40(1):561–592, 2012.

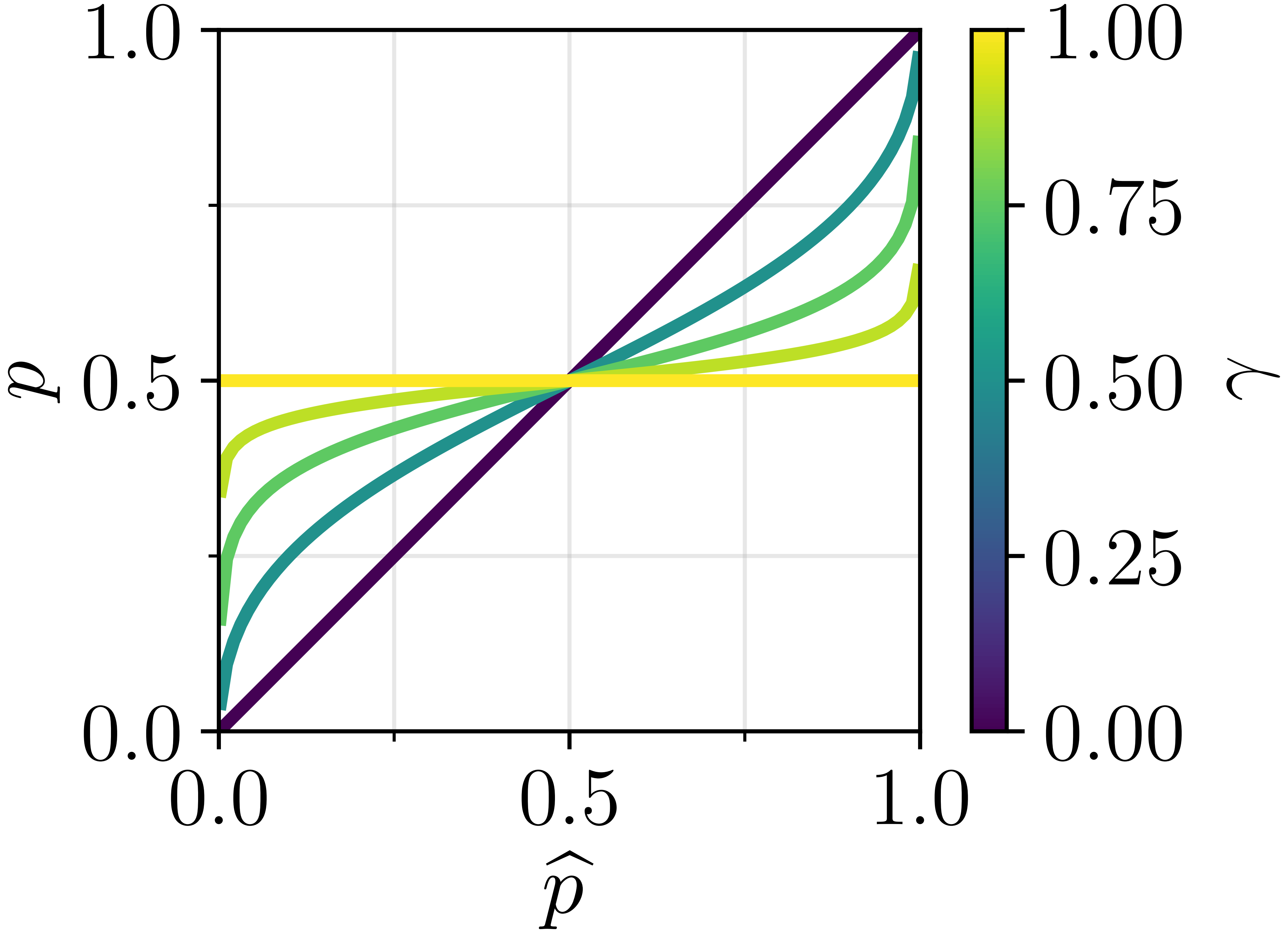
Can we recover correct probability estimate?

- Theoretical minimizer of $L(\boldsymbol{p}, \widehat{\boldsymbol{p}})$ is distorted
- **Idea: We can recover by applying the inverse of distortion**

Theoretical minimizer of focal loss



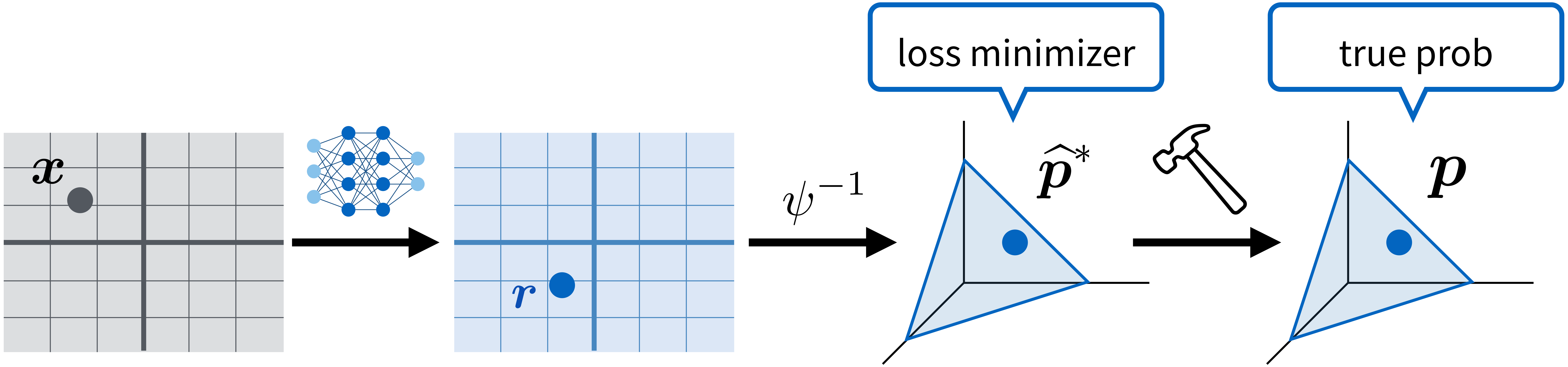
Theoretical minimizer of generalized CE loss



Inverse of “distortion”

Theorem Assume loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is local with each component $\ell_y(\widehat{p}_y)$, and continuously differentiable and invertible. Then, the conditional risk $L(\boldsymbol{p}, \cdot)$ has a minimizer $\widehat{\boldsymbol{p}}^*$ satisfying

$$p_y = \frac{[\ell_y'(\widehat{p}_y^*)]^{-1}}{\sum_{i=1}^K [\ell_y'(\widehat{p}_i^*)]^{-1}}$$



Inverse of “distortion”

Theorem Assume loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is local with each component $\ell_y(\widehat{p}_y)$, and continuously differentiable and invertible. Then, the conditional risk $L(\boldsymbol{p}, \cdot)$ has a minimizer $\widehat{\boldsymbol{p}}^*$ satisfying

$$p_y = \frac{[\ell'_y(\widehat{p}^*_y)]^{-1}}{\sum_{i=1}^K [\ell'_y(\widehat{p}^*_i)]^{-1}}$$

- Define Ψ-transform by

$$\Psi(\widehat{\boldsymbol{p}})_y = \frac{[\ell'_y(\widehat{p}^*_y)]^{-1}}{\sum_{i=1}^K [\ell'_y(\widehat{p}^*_i)]^{-1}}$$

- **Proof sketch**: solve the KKT condition for the Lagrangian

$$\mathcal{L}(\widehat{\boldsymbol{p}}, \beta) = \sum_{y=1}^K p_y \ell_y(\widehat{p}_y) + \beta\left(\sum_{y=1}^K \widehat{p}_y - 1\right)$$

Calm composite loss

- Define Ψ-transform by

$$\Psi(\widehat{\boldsymbol{p}})_y = \frac{[\ell'_y(\widehat{p}^*_y)]^{-1}}{\sum_{i=1}^K [\ell'_y(\widehat{p}^*_i)]^{-1}}$$

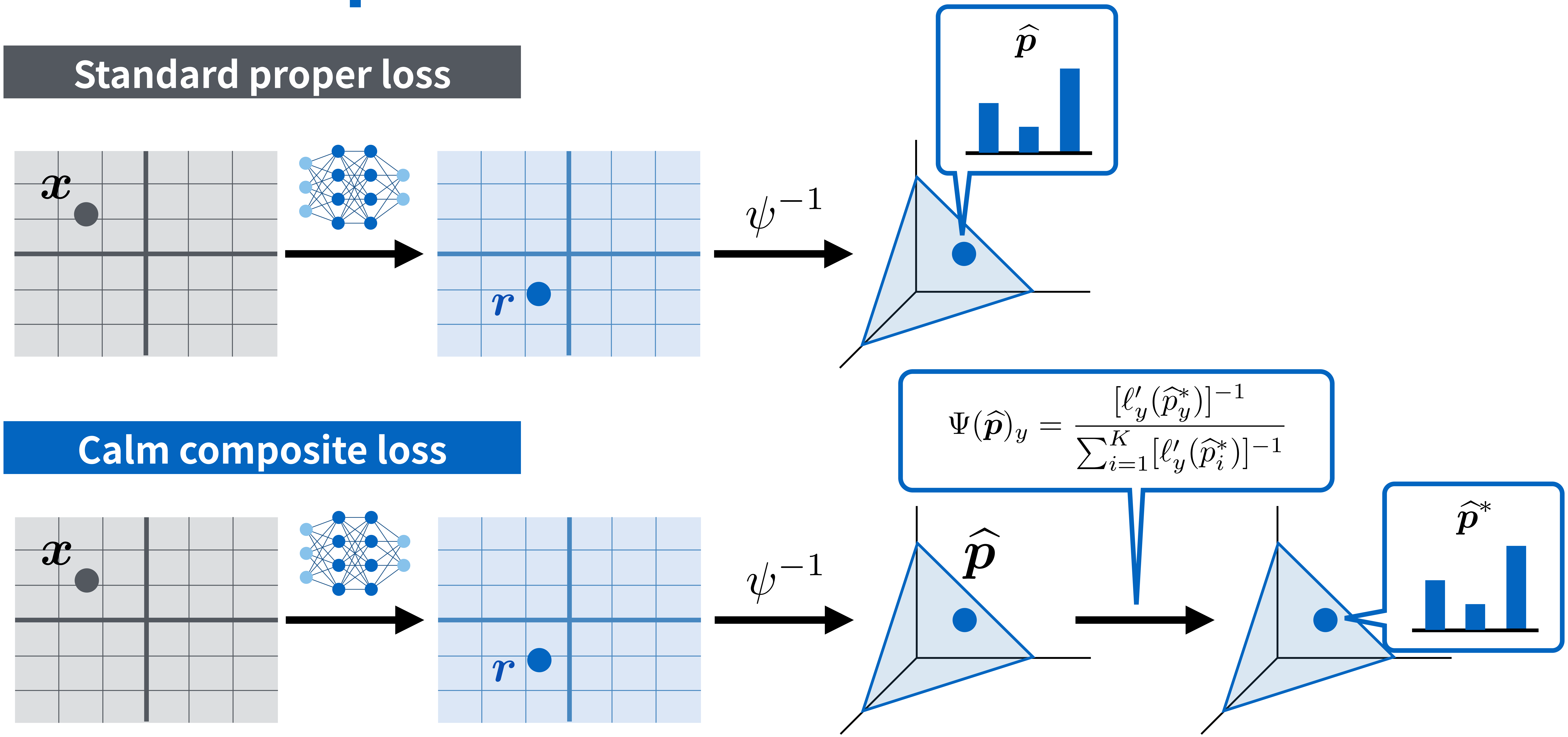
- **Q** When is Ψ bijective?

Theorem Assume loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is local with each component $\ell_y(\widehat{p}_y)$, and continuously twice differentiable and invertible. Then, Ψ is bijective if for $y \in [K]$, the following conditions hold:

$$\ell'_y < 0, \quad \ell''_y > 0, \quad \text{and} \quad \lim_{p \downarrow 0} \ell'_y(p) = -\infty$$

calmness condition

Calm composite loss



Standard proper loss

Calm composite loss

$$\Psi(\widehat{\boldsymbol{p}})_y = \frac{[\ell'_y(\widehat{p}^*_y)]^{-1}}{\sum_{i=1}^{K}[\ell'_y(\widehat{p}^*_i)]^{-1}}$$

Calmness condition

Theorem Assume loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is local with each component $\ell_y(\widehat{p}_y)$, and continuously differentiable and invertible. Then, the conditional risk $L(\boldsymbol{p}, \cdot)$ has a minimizer $\widehat{\boldsymbol{p}}^*$ satisfying

$$p_y = \frac{[\ell'_y(\widehat{p}^*_y)]^{-1}}{\sum_{i=1}^K [\ell'_y(\widehat{p}^*_i)]^{-1}}$$

- **Example.** Generalized cross-entropy loss $\ell_y(\widehat{\boldsymbol{p}}) = (1 - \widehat{p}_y^\gamma)/\gamma$
  - $\ell'_y(p) = -\frac{1}{p^{1-\gamma}} < 0$
  - $\lim_{p \downarrow 0} \ell'_y(p) = -\infty$
  - $\ell''_y(p) = \frac{1-\gamma}{p^{2-\gamma}} > 0$

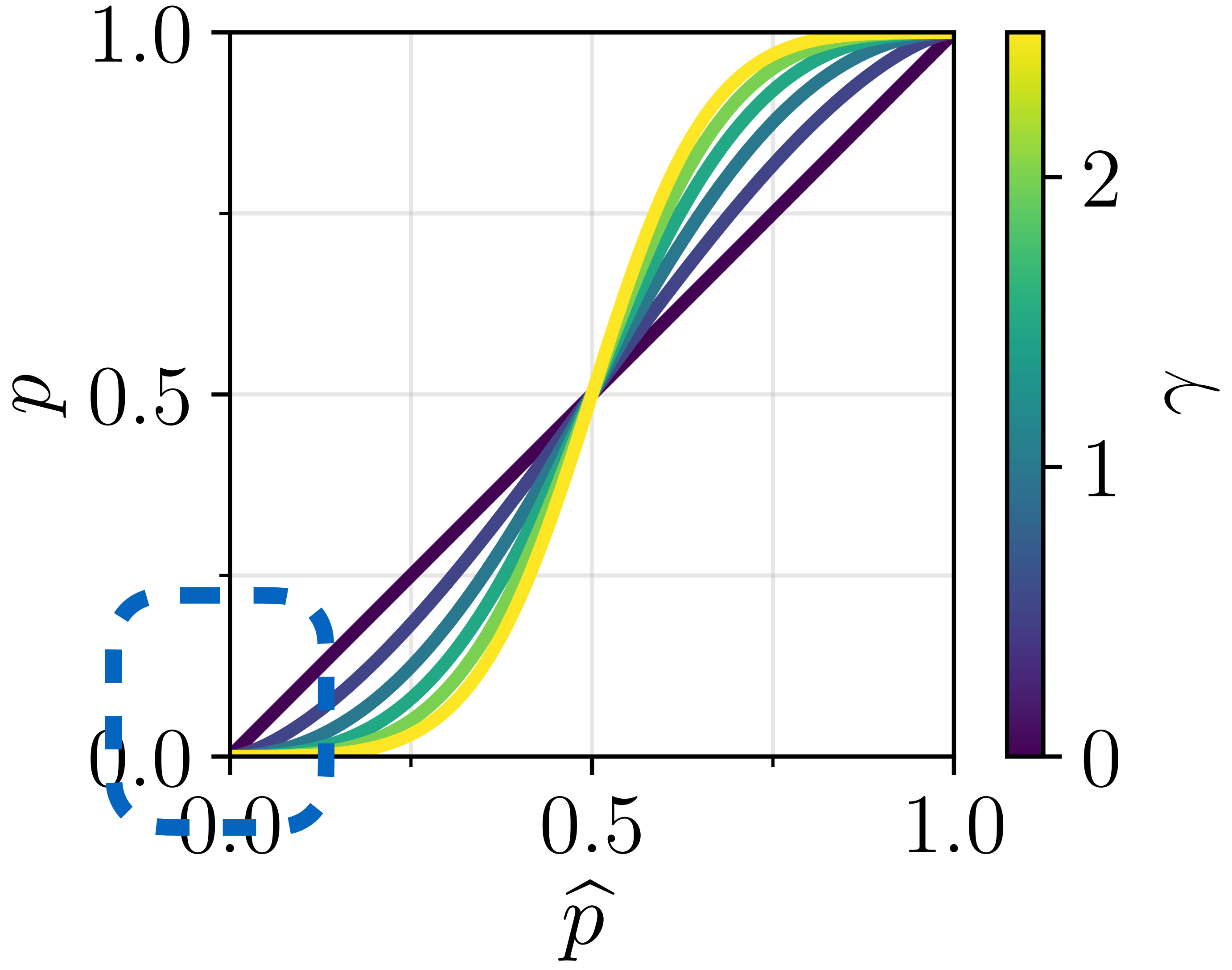
Calmness condition

Theorem Assume loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is local with each component $\ell_y(\widehat{p}_y)$, and continuously differentiable and invertible. Then, the conditional risk $L(\boldsymbol{p}, \cdot)$ has a minimizer $\widehat{\boldsymbol{p}}^*$ satisfying

$$p_y = \boxed{\frac{[\ell'_y(\widehat{p}^*_y)]^{-1}}{\sum_{i=1}^K [\ell'_y(\widehat{p}^*_i)]^{-1}}}$$

ensuring surjectivity

Theoretical minimizer of focal loss (**calm**)



$$p_y = \frac{[\ell'_y(\widehat{p}^*_y)]^{-1}}{\sum_{i=1}^K [\ell'_y(\widehat{p}^*_i)]^{-1}}$$

If $\ell'_y(0)$ is finite, p_y cannot go to 0

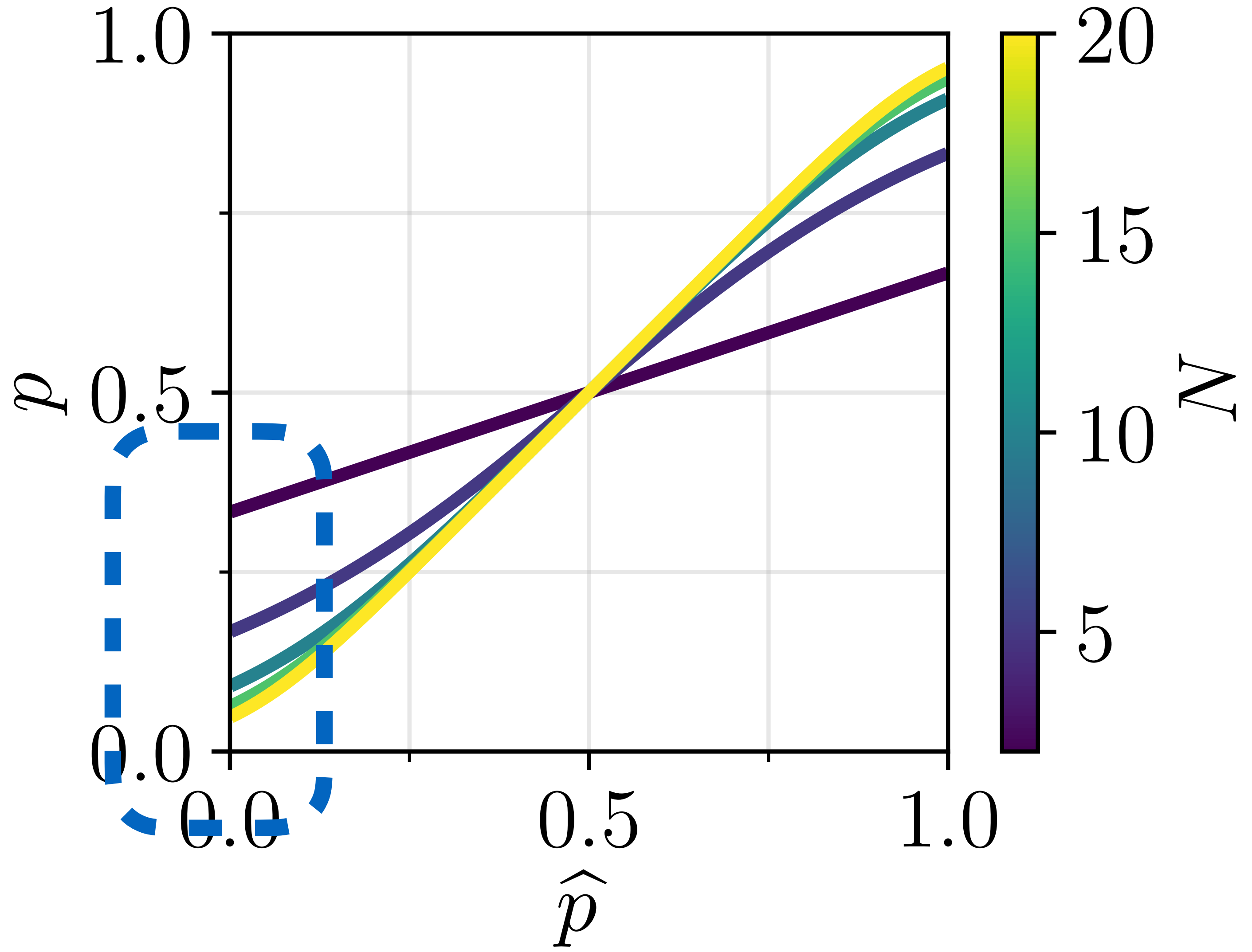
Calmness condition

Theorem Assume loss $\boldsymbol{\ell} : \triangle^K \to \mathbb{R}^K$ is local with each component $\ell_y(\widehat{p}_y)$, and continuously differentiable and invertible. Then, the conditional risk $L(\boldsymbol{p}, \cdot)$ has a minimizer $\widehat{\boldsymbol{p}}^*$ satisfying

$$p_y = \frac{[\ell'_y(\widehat{p}^*_y)]^{-1}}{\sum_{i=1}^K [\ell'_y(\widehat{p}^*_i)]^{-1}}$$

ensuring surjectivity

Theoretical minimizer of Taylor CE loss (**not calm**)



$$p_y = \frac{[\ell'_y(\widehat{p}^*_y)]^{-1}}{\sum_{i=1}^K [\ell'_y(\widehat{p}^*_i)]^{-1}}$$

If $\ell'_y(0)$ is finite, p_y cannot go to 0

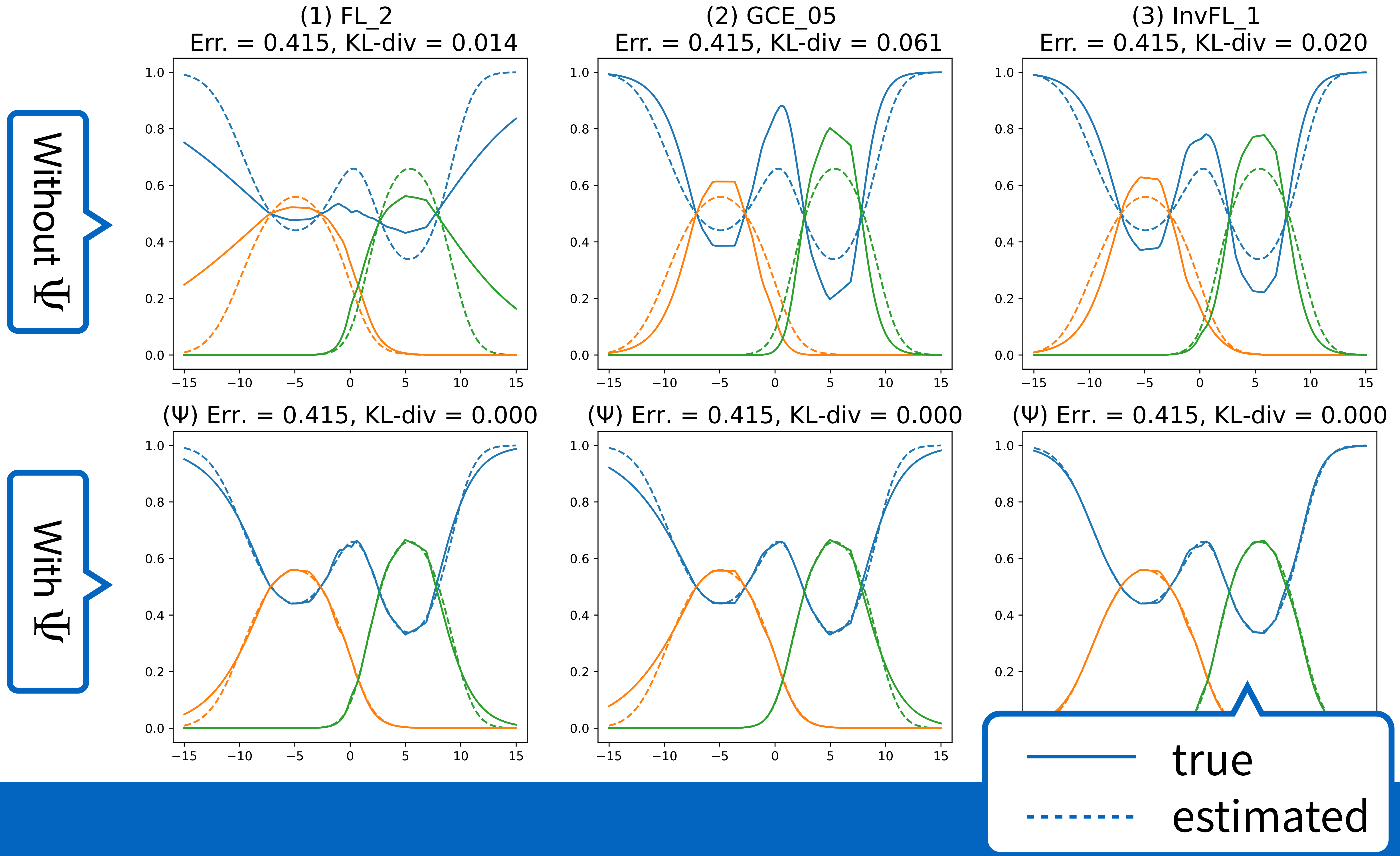
List of loss functions

Loss	$\ell(q)$	Proper	Calm
Log	$-\log q$	☑	☑
Focal	$-(1-q)^\gamma \log q$	☒	☑
Inverse focal $\gamma \in [0,1]$	$-(1+q)^\gamma \log q$	☒	☑
Generalized CE	$(1-q^\gamma)/\gamma$	☒	☑
MAE	$1-q$	☒	☒
Power	$(1-q)^\gamma$	☒	☒

⬆ all of them are local losses

Numerical simulation (calm losses)

- Data: 3-class classification with each 1D Gaussian
- Model: 3-layer MLP



(1) FL_2
Err. = 0.415, KL-div = 0.014

(2) GCE_05
Err. = 0.415, KL-div = 0.061

(3) InvFL_1
Err. = 0.415, KL-div = 0.020

(Ψ) Err. = 0.415, KL-div = 0.000

(Ψ) Err. = 0.415, KL-div = 0.000

(Ψ) Err. = 0.415, KL-div = 0.000

Without Ψ

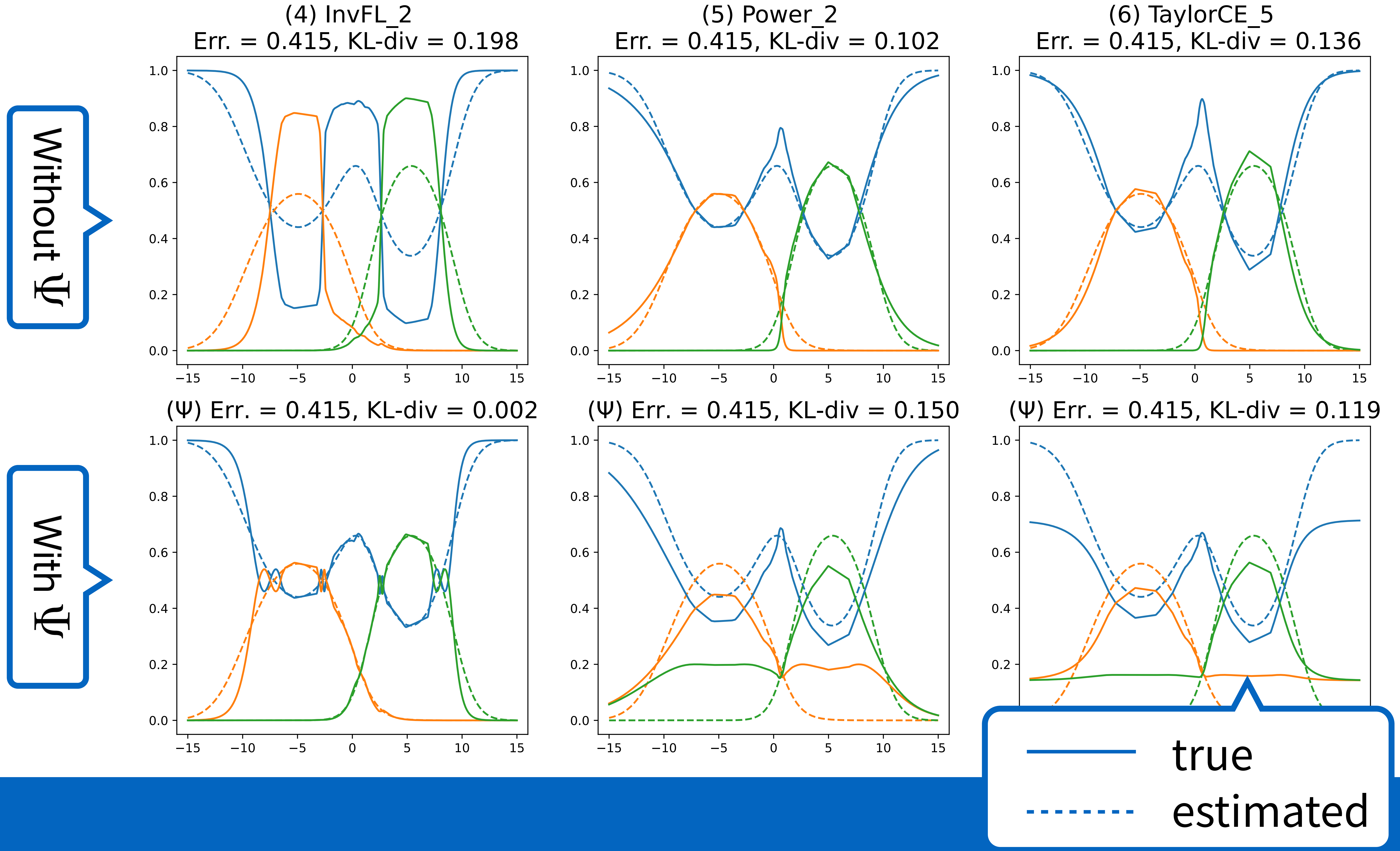
With Ψ

true

estimated

Numerical simulation (non-calm losses)

- Data: 3-class classification with each 1D Gaussian
- Model: 3-layer MLP



Without Ψ

With Ψ

(4) InvFL_2
Err. = 0.415, KL-div = 0.198

(5) Power_2
Err. = 0.415, KL-div = 0.102

(6) TaylorCE_5
Err. = 0.415, KL-div = 0.136

(Ψ) Err. = 0.415, KL-div = 0.002

(Ψ) Err. = 0.415, KL-div = 0.150

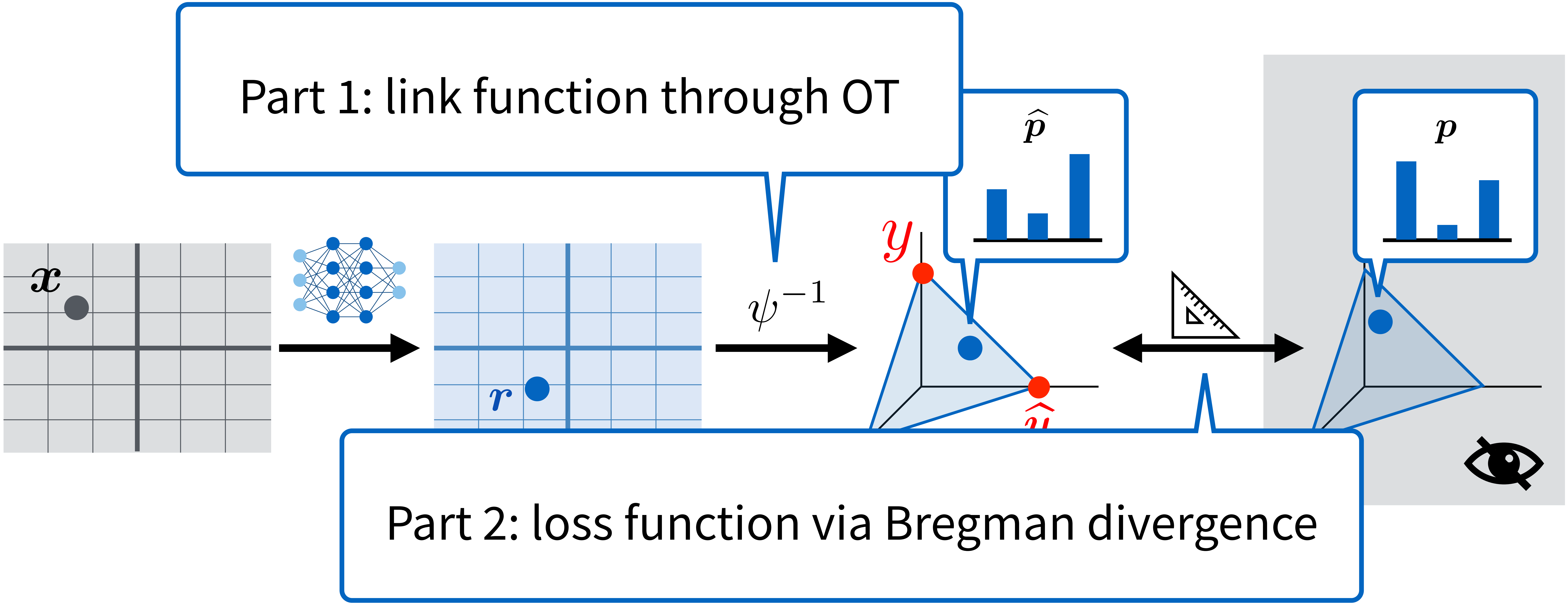
(Ψ) Err. = 0.415, KL-div = 0.119

true

estimated

Summary

Convex analysis enriches calibration



Part 1: link function through OT

Part 2: loss function via Bregman divergence