

Learning Energy Networks with Generalized Fenchel-Young Losses



Mathieu Blondel



Felipe Llinares



Robert Dadashi



Léonard Hussenot

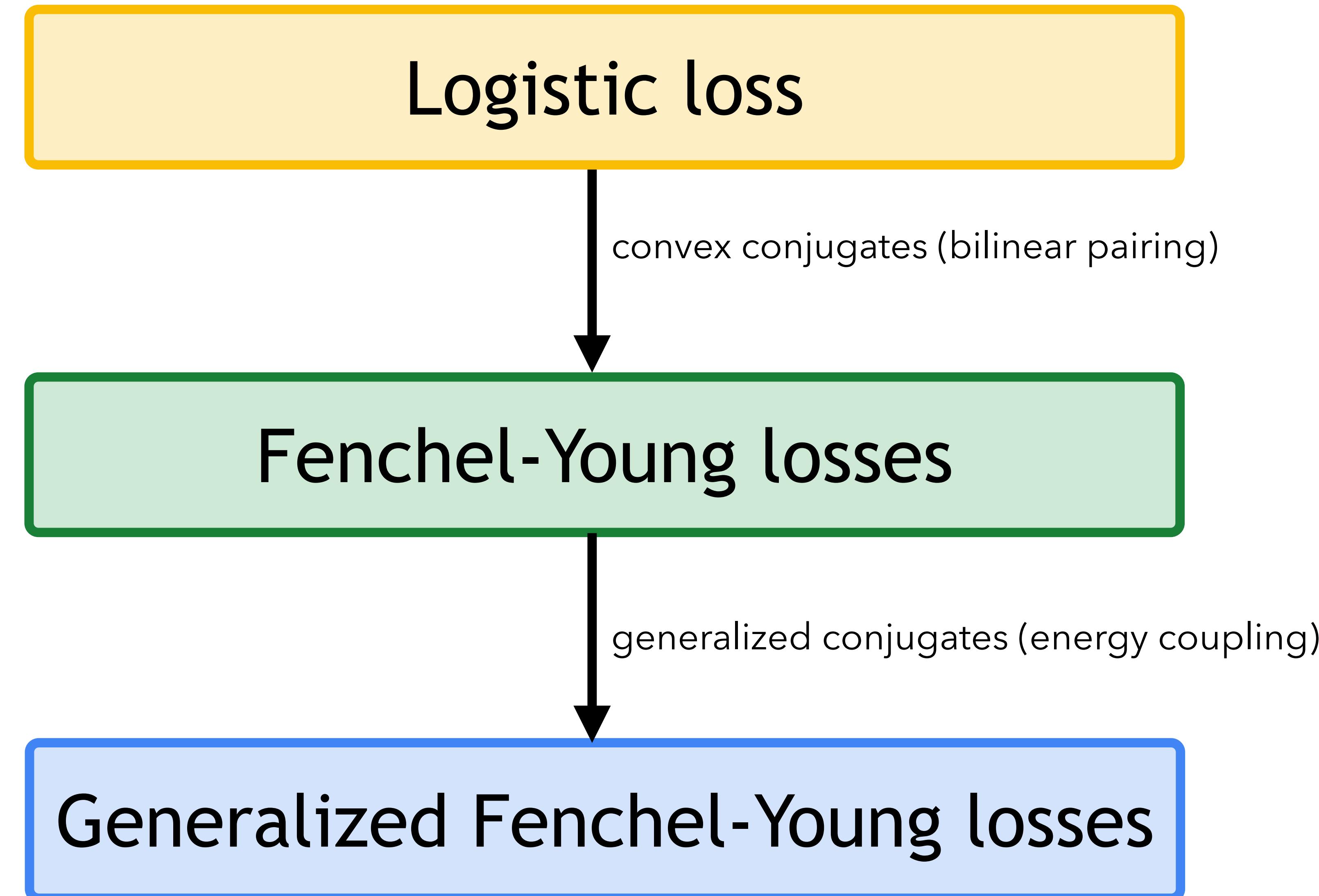


Matthieu Geist

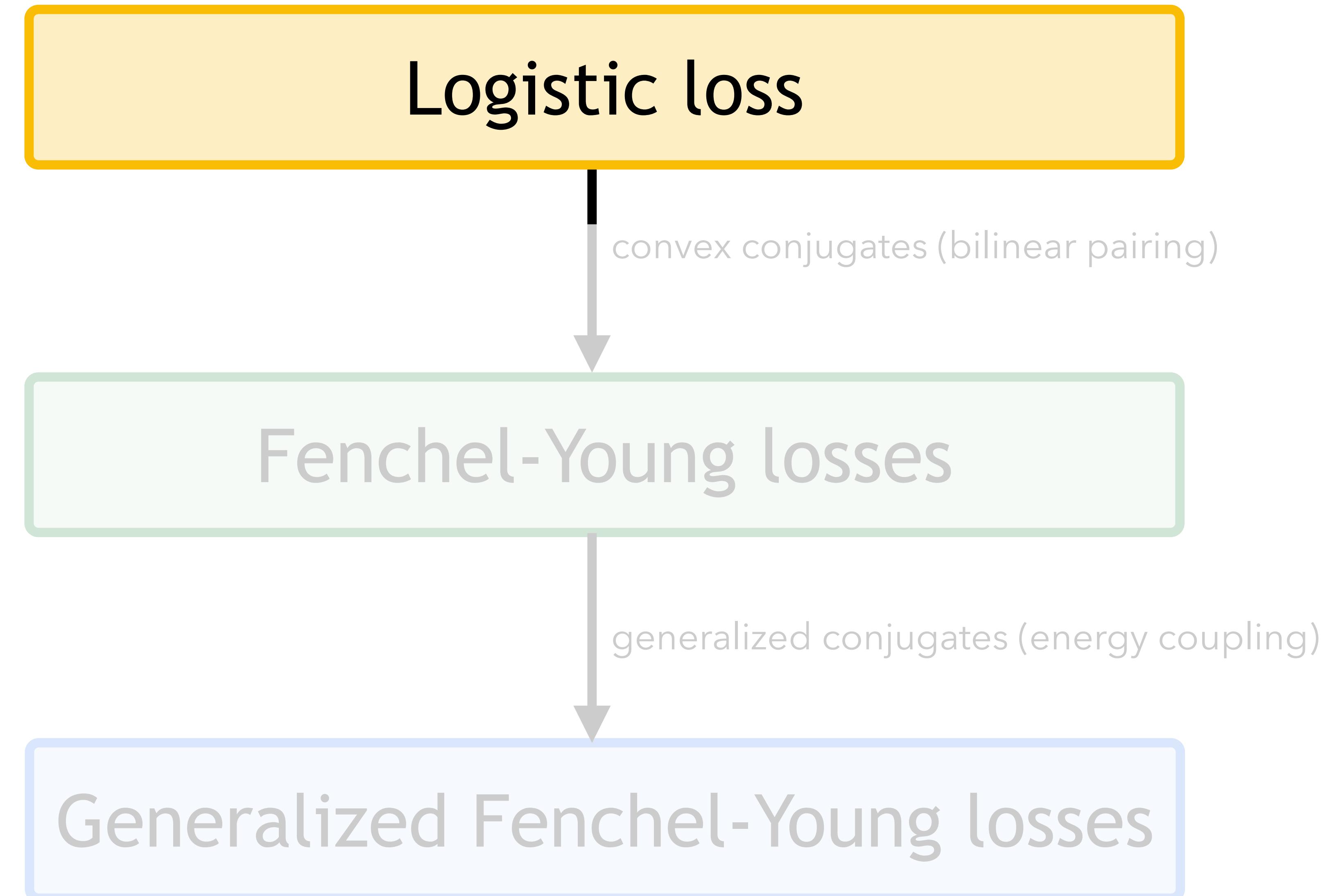
PGMODAYS
November 28th, 2023

 Google DeepMind

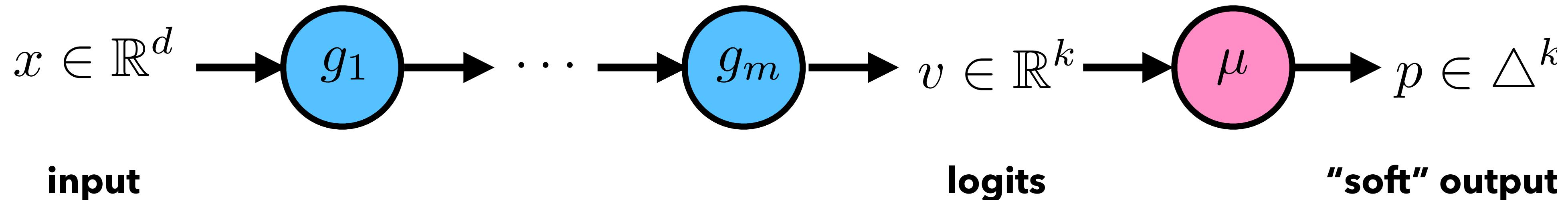
Outline



Outline



Neural networks with softmax output layer



$$v = g(x) = (g_m \circ \dots \circ g_1)(x) \in \mathbb{R}^k \quad \text{feature layers}$$

$$p = \mu(v) = \frac{\exp(v)}{\sum_{j=1}^k \exp(v_j)} \in \Delta^k \quad \text{softmax output layer}$$

Goal: learn g such that $\mu(v) \approx y$ for all (x, y) pairs

Logistic loss

$$L_{\log}(v, y) = \text{logsumexp}(v) - \langle v, y \rangle$$

$$= \log \sum_{j=1}^k \exp(v_j) - \langle v, y \rangle$$

logits $v = g(x)$

ground-truth label $y \in \{e_1, \dots, e_k\}$

Logistic loss gradient

$$\begin{aligned}\nabla_1 L_{\log}(v, y) &= \mu(v) - y \\ &= \mathbb{E}_{Y \sim p}[Y] - y\end{aligned}$$

softmax

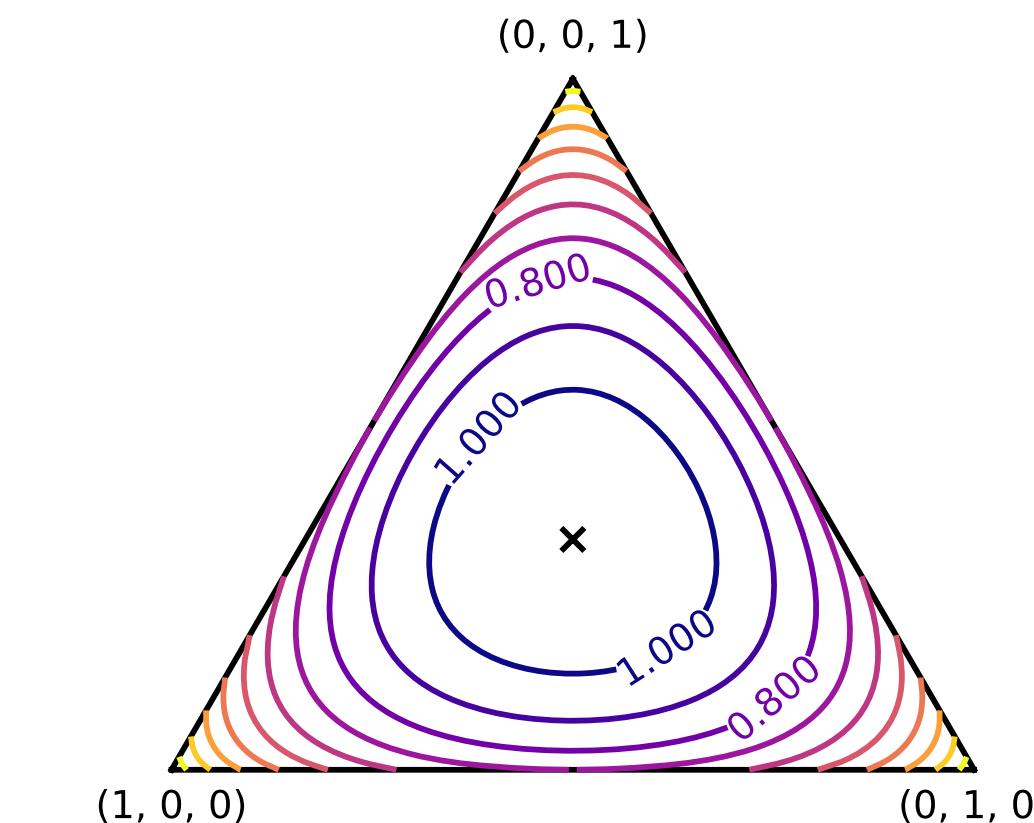
$$p = \mu(v) = \frac{\exp(v)}{\sum_{j=1}^k \exp(v_j)} \in \Delta^k$$

ground-truth label

$$y \in \{e_1, \dots, e_k\}$$

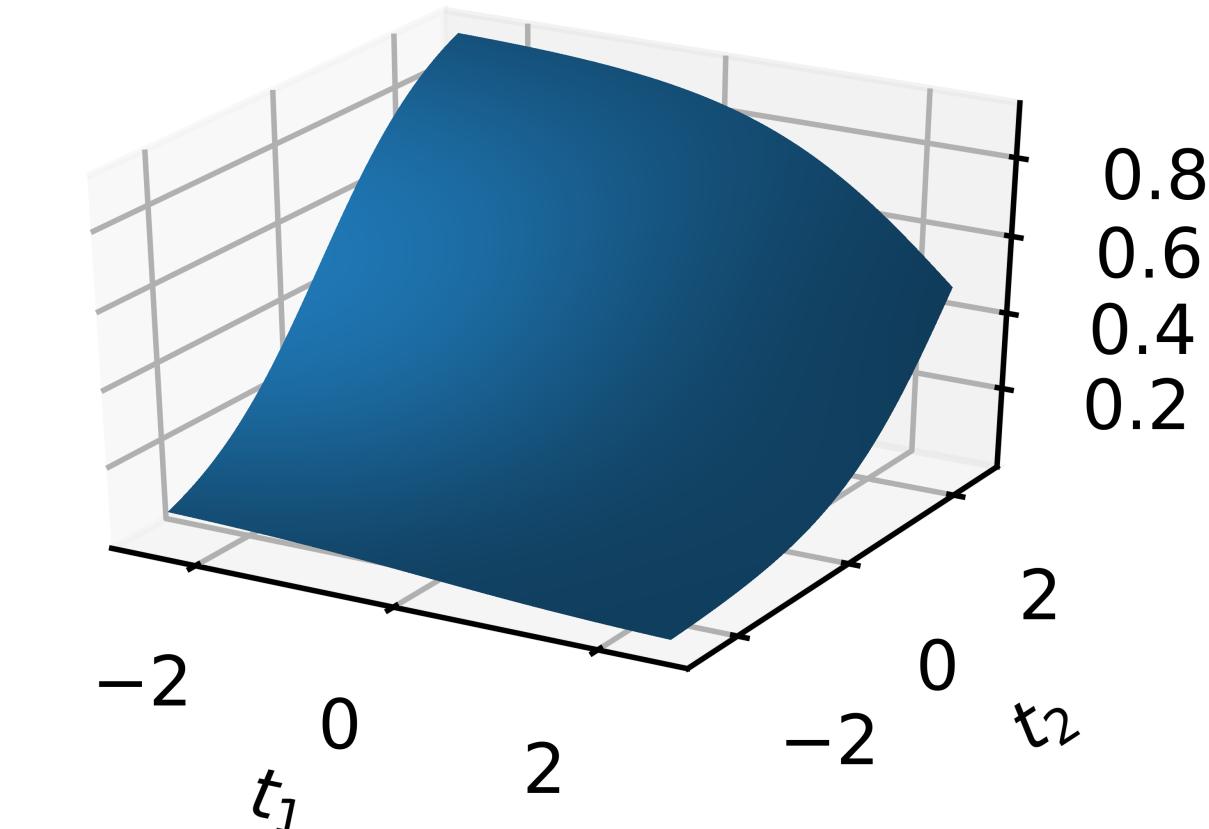
Softmax as an argmax output layer

$$\begin{aligned}\text{logsumexp}(v) &= \log \sum_{j=1}^k \exp(v_j) \\ &= \max_{p \in \Delta^k} \langle v, p \rangle - \langle p, \log p \rangle\end{aligned}$$



Contours of Shannon's entropy

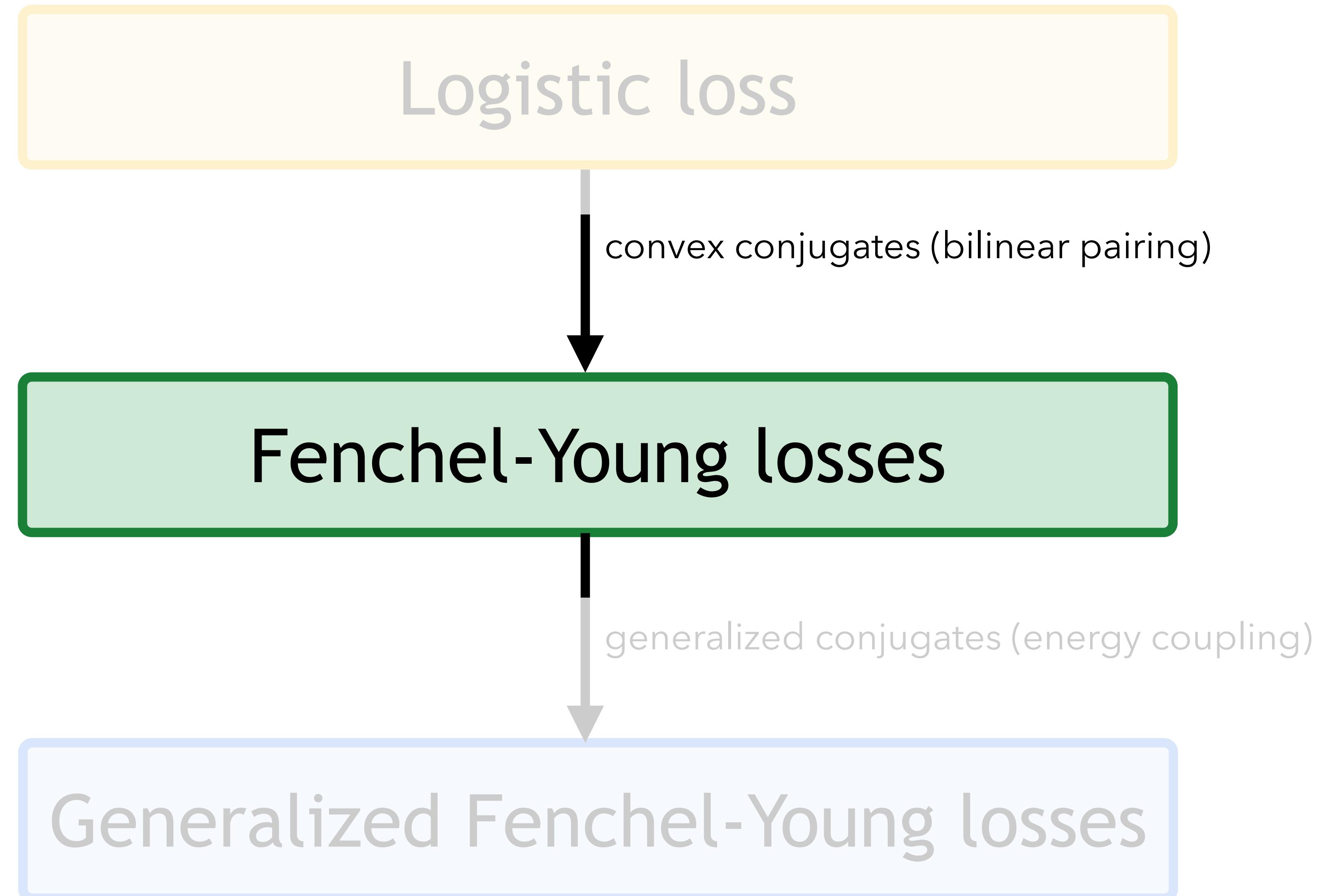
$$\begin{aligned}\mu(v) &= \frac{\exp(v)}{\sum_{j=1}^k \exp(v_j)} \\ &= \operatorname{argmax}_{p \in \Delta^k} \langle v, p \rangle - \langle p, \log p \rangle \\ &= \nabla \text{logsumexp}(v)\end{aligned}$$



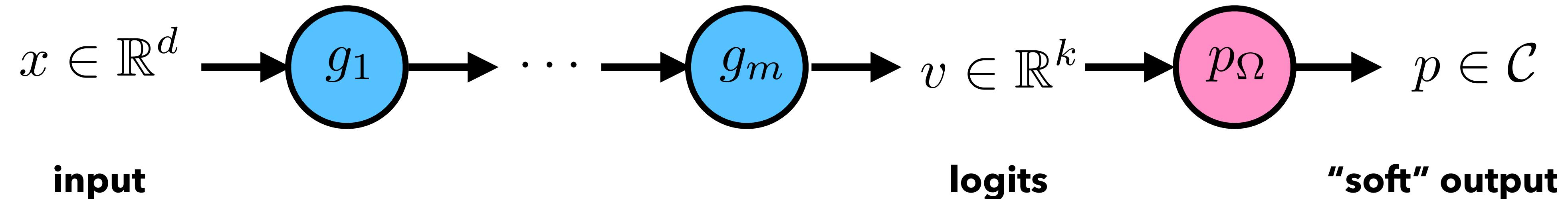
Surface of the softmax

$$\mu(t_1, t_2, 0)$$

Outline



Regularized argmax output layers



$$p = p_\Omega(v) = \underset{p \in \mathcal{C}}{\operatorname{argmax}} \langle v, p \rangle - \Omega(p)$$

Goal: learn g such that $p_\Omega(v) \approx y$ for all (x, y) pairs

Convex conjugate functions

$$\Omega^*(v) = \max_{p \in \mathcal{C}} \langle v, p \rangle - \Omega(p)$$

$$\Omega: \mathcal{C} \rightarrow \mathbb{R}$$

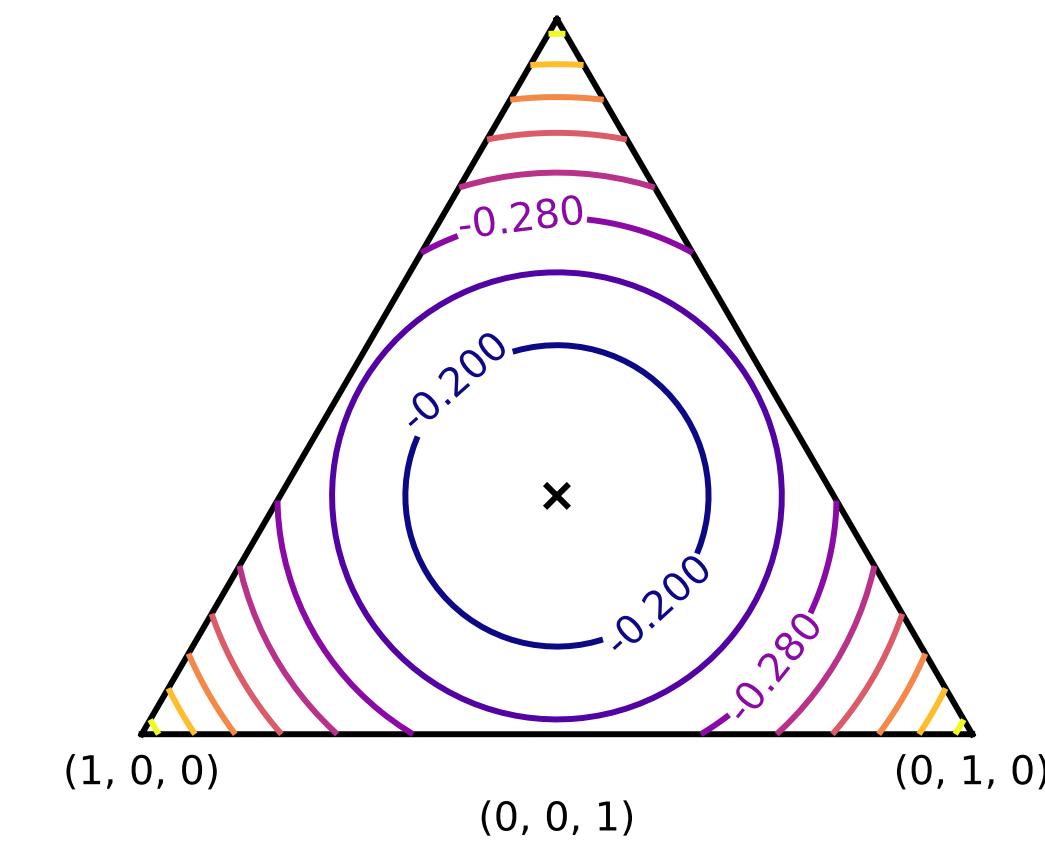
$$p_\Omega(v) = \operatorname{argmax}_{p \in \mathcal{C}} \langle v, p \rangle - \Omega(p)$$

$$= \nabla \Omega^*(v)$$

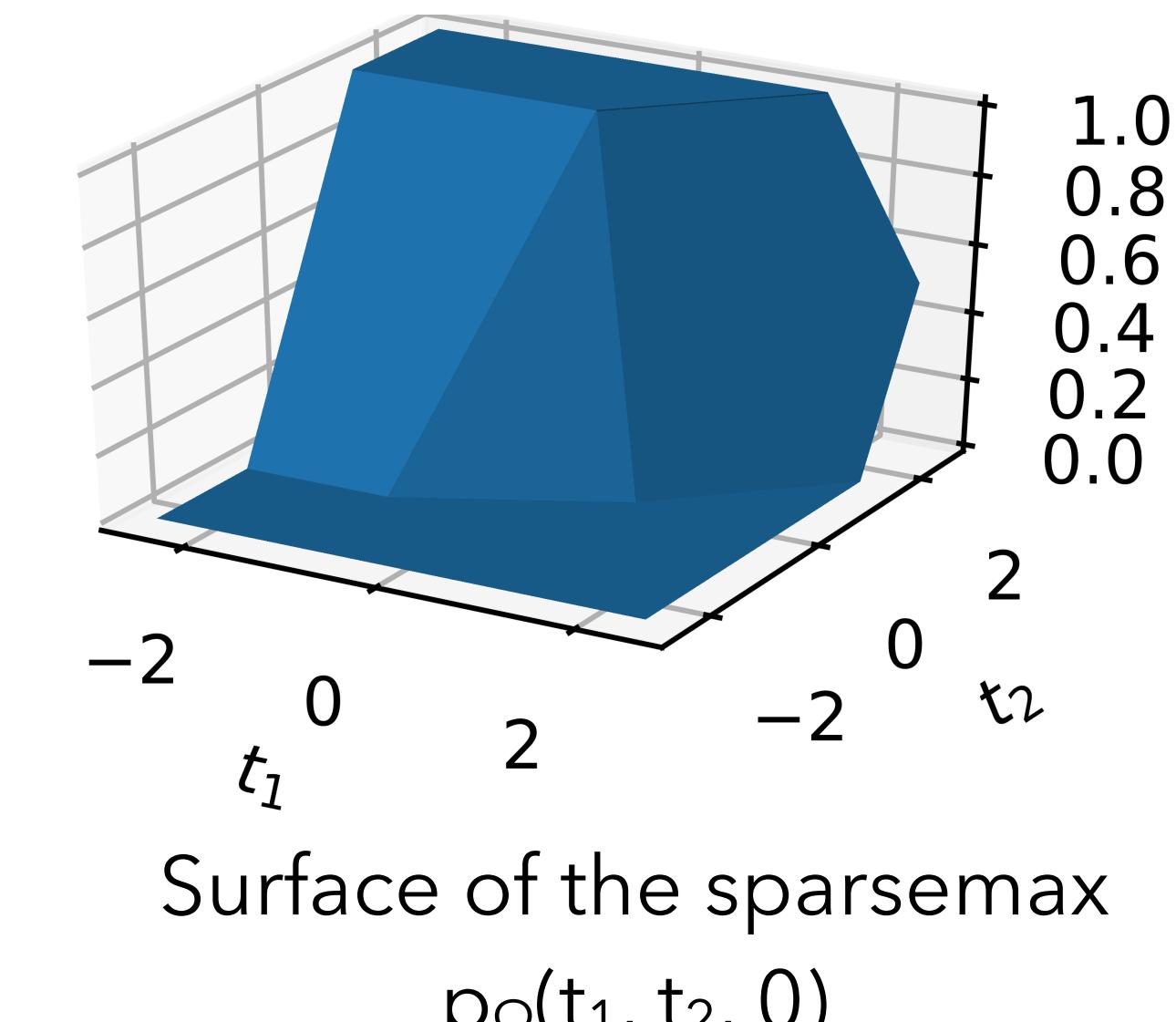
$f(v)$ is closed and convex

\Updownarrow

$$\exists \Omega \text{ s.t. } f(v) = \Omega^*(v)$$



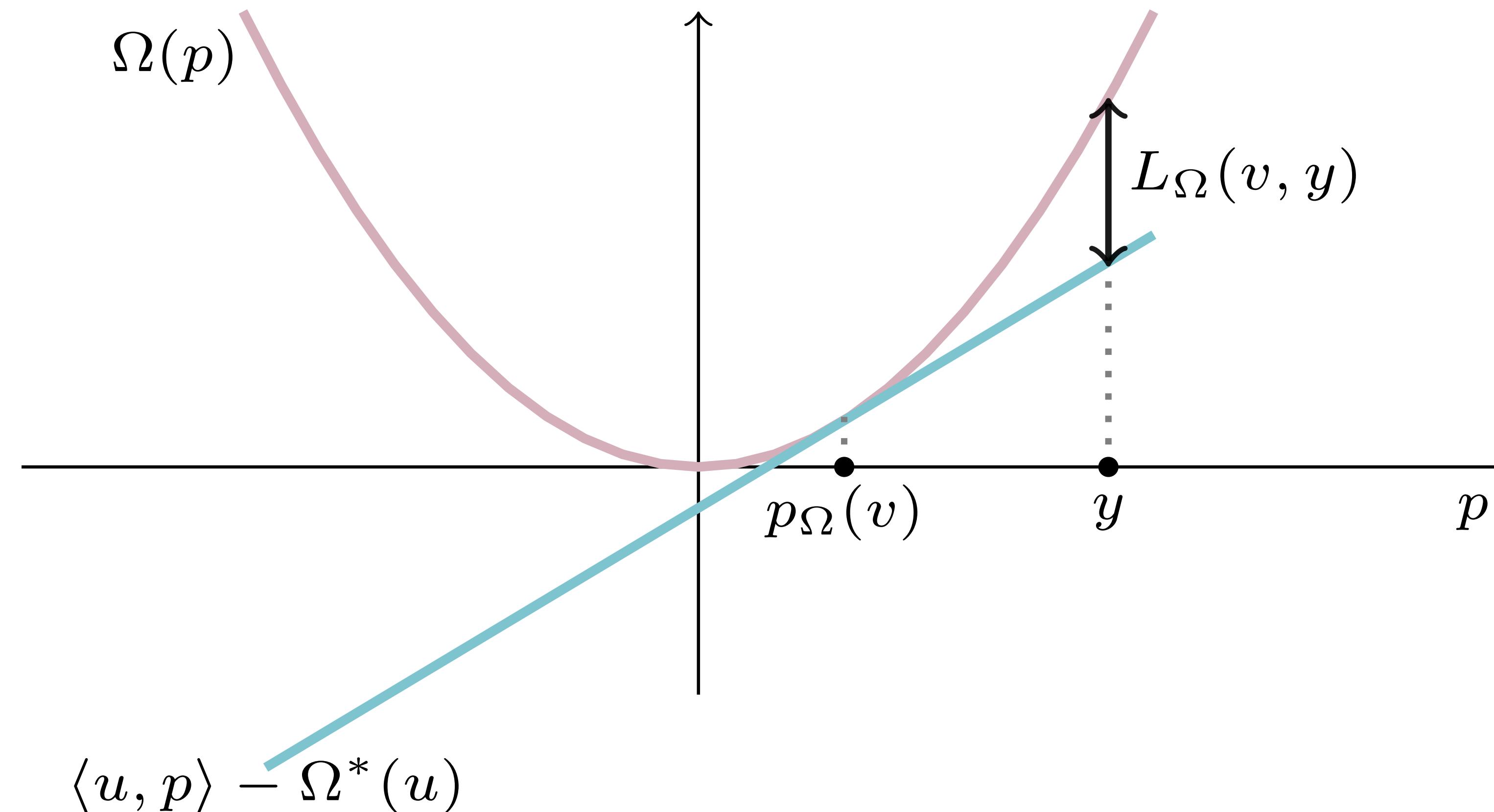
$$\Omega(p) = \frac{1}{2} \langle p, 1 - p \rangle$$



Fenchel-Young loss functions

(Blondel, Martins, Niculae, 2020)

$$L_{\Omega}(v, y) := \Omega^*(v) + \Omega(y) - \langle v, y \rangle$$



Fenchel-Young loss properties

Non-negativity

$$L_{\Omega}(v, y) \geq 0$$

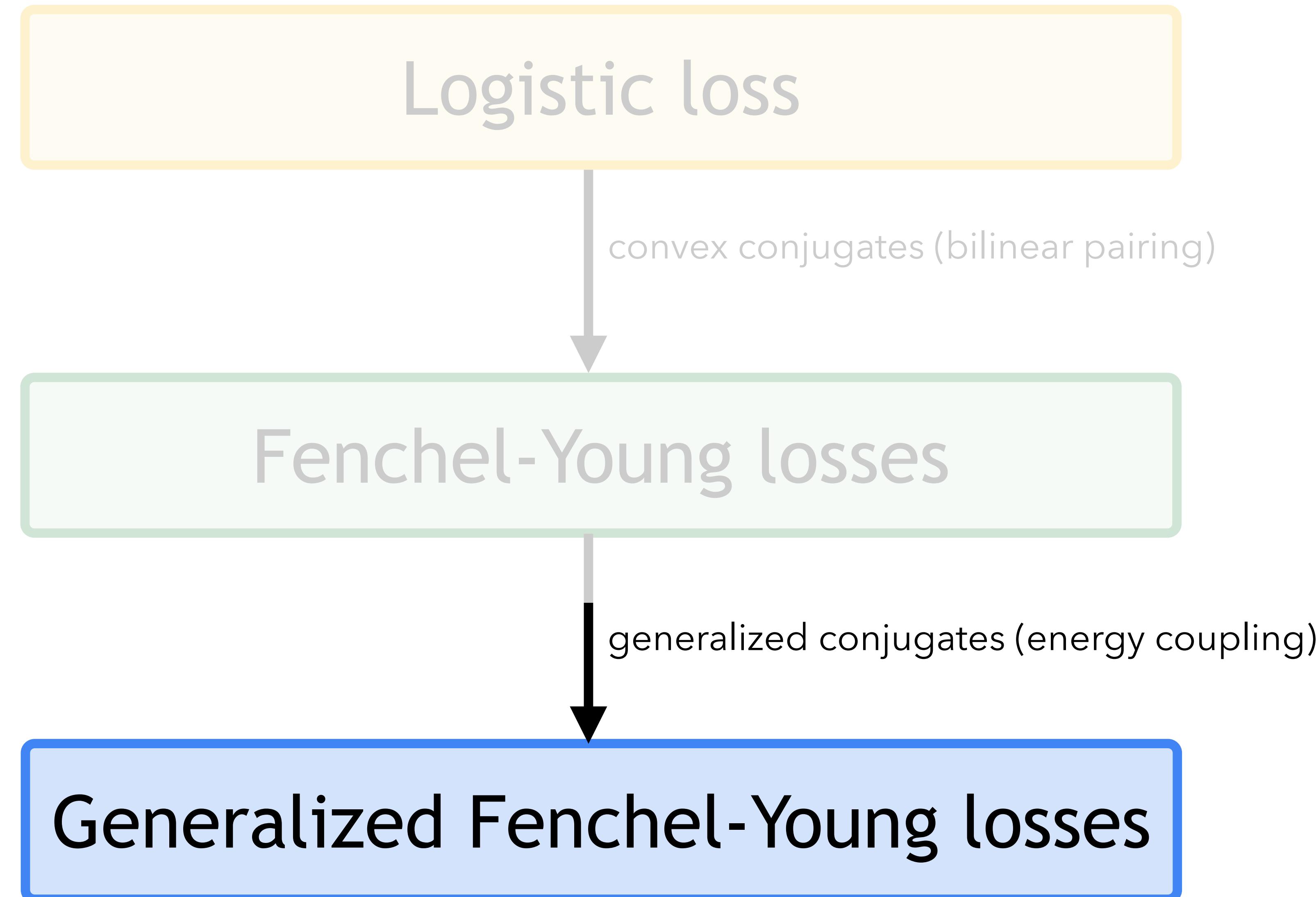
Zero loss

$$L_{\Omega}(v, y) = 0 \Leftrightarrow p_{\Omega}(v) = y$$

Gradient

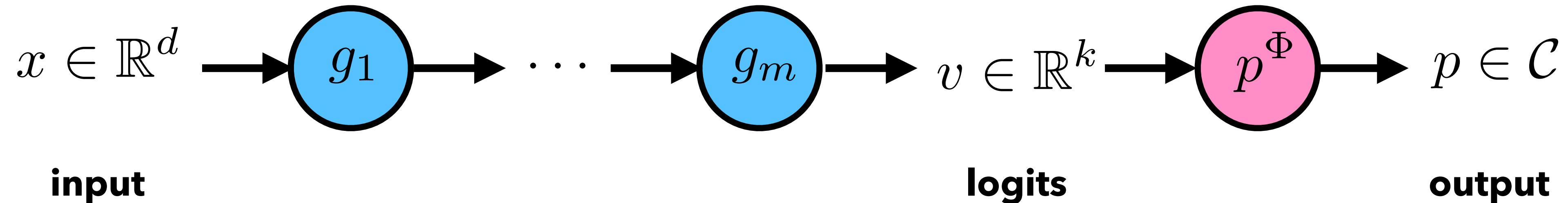
$$\begin{aligned}\nabla_1 L_{\Omega}(v, p) &= \nabla \Omega^*(v) - y \\ &= p_{\Omega}(v) - y\end{aligned}$$

Outline



Energy networks

(LeCun, 2006)



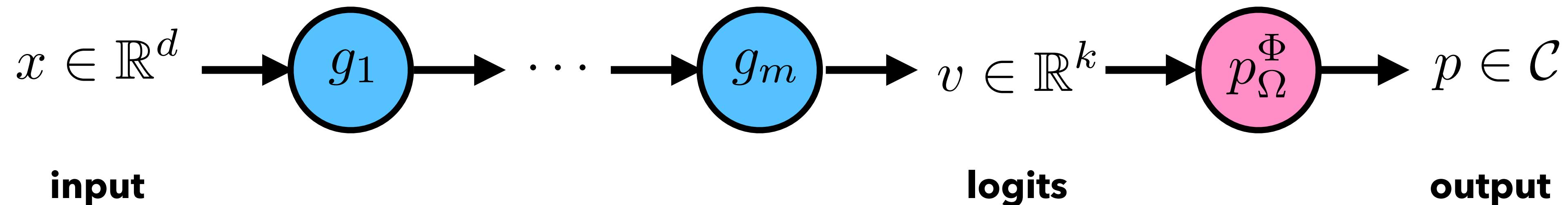
$$p = p^\Phi(v) = \operatorname{argmax}_{p \in \mathcal{C}} \Phi(v, p) \in \mathcal{C}$$

**output maximizing
an energy Φ**

Goal: learn g such that $p^\Phi(v) \approx y$ for all (x, y) pairs

Regularized energy networks

(Blondel et al, 2022)



$$p = p_{\Omega}^{\Phi}(v) = \operatorname{argmax}_{p \in \mathcal{C}} \Phi(v, p) - \Omega(p) \in \mathcal{C}$$

**output maximizing
a regularized
energy**

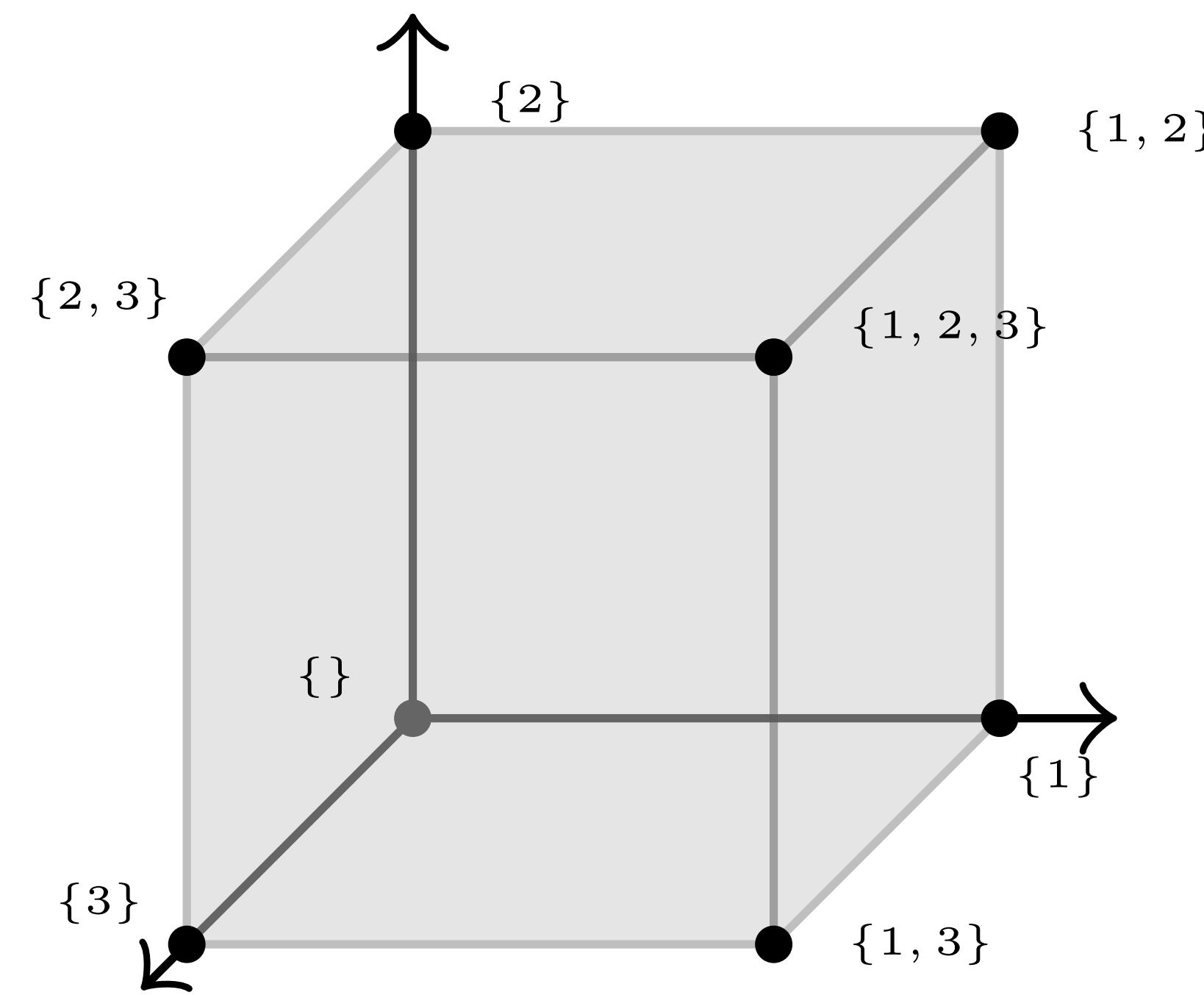
$$\Phi: \mathcal{V} \times \mathcal{C} \rightarrow \mathbb{R} \quad \Omega: \mathcal{C} \rightarrow \mathbb{R}$$

Example: linear-quadratic energy

(Blondel et al, 2022)

$$p_{\Omega}^{\Phi}(v) = \operatorname{argmax}_{p \in [0,1]^k} \langle u, p \rangle + \frac{1}{2} \langle p, Up \rangle - \Omega(p)$$

$v = (u, U)$ u_j : weight of label j $U_{i,j}$: weight of labels i and j



Regularized energy networks

(Blondel et al, 2022)

	$\Phi(v, p)$	v	p
GLM	$\langle v, p \rangle$	linear	linear
Linear-quadratic	$\frac{1}{2} \langle p, Ap \rangle + \langle p, b \rangle$	linear	quadratic
Rectifier network	$\langle \text{relu}(v), Up \rangle$	convex	linear
Maxout network	$p \cdot \max(v)$	convex	linear
LSE network	$p \cdot \text{LSE}^\gamma(v)$	convex	linear
ICNN	$-\text{ICNN}(v, p)$	nonconvex	concave
Probabilistic	$\sum_{y \in \mathcal{Y}} p(y) E(v, y)$	nonconvex	linear
Arbitrary	$\Phi(v, p)$	nonconvex	nonconcave

Generalized conjugates

$$\Phi: \mathcal{V} \times \mathcal{C} \rightarrow \mathbb{R} \quad \Omega: \mathcal{C} \rightarrow \mathbb{R}$$

(Moreau, 1966)

$$\Omega^\Phi(v) := \max_{p \in \mathcal{C}} \Phi(v, p) - \Omega(p)$$

$$p_\Omega^\Phi(v) := \operatorname{argmax}_{p \in \mathcal{C}} \Phi(v, p) - \Omega(p)$$

$F(v)$ is Φ -convex



$$\exists \Omega \text{ s.t. } F(v) = \Omega^\Phi(v)$$

Generalized conjugate properties

1. **Generalized Fenchel-Young inequality:** for all $v \in \mathcal{V}$ and $p \in \mathcal{C}$,

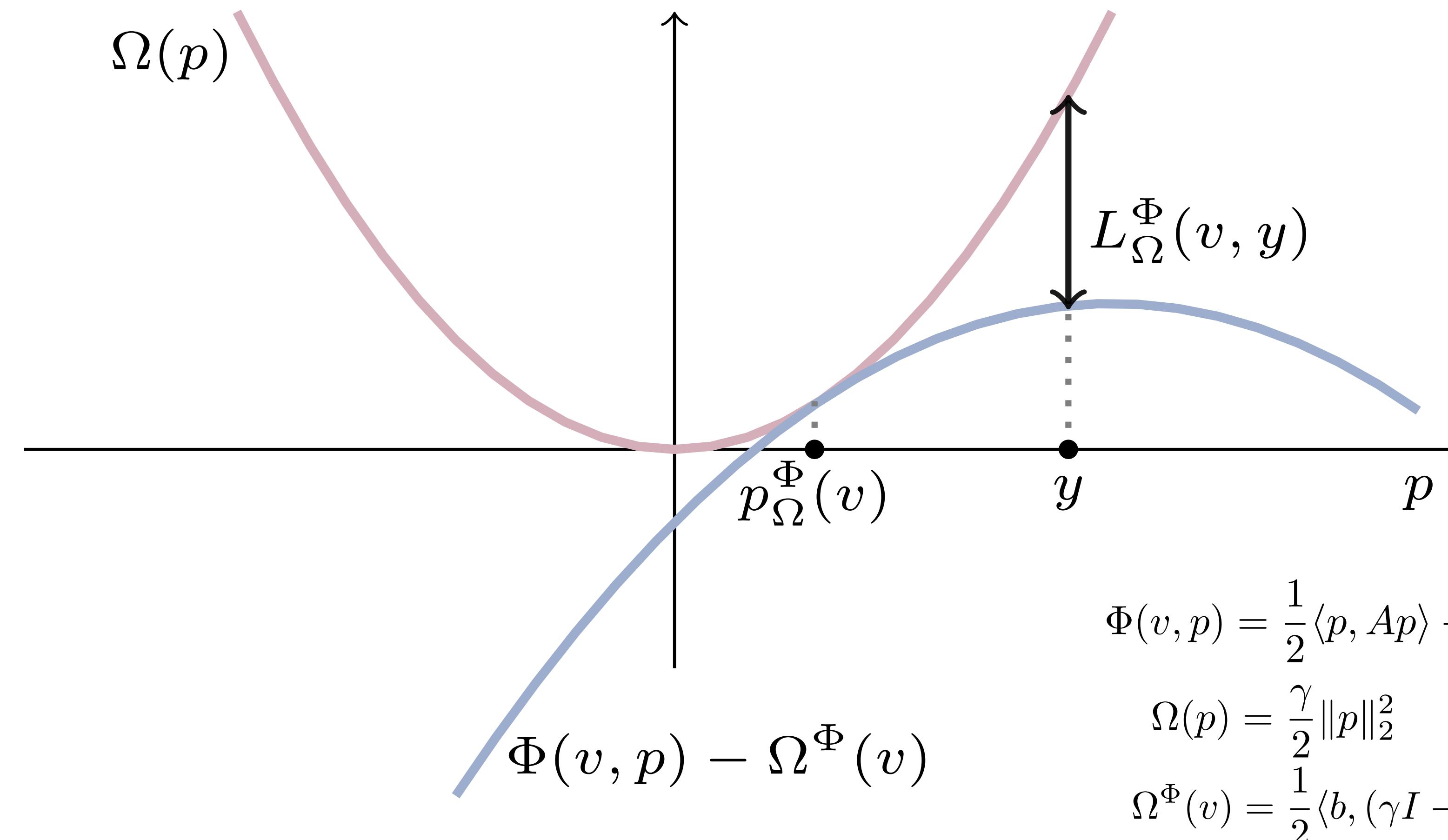
$$\Omega^\Phi(v) + \Omega(p) - \Phi(v, p) \geq 0.$$

2. **Convexity:** If $\Phi(v, p)$ is convex in v , then $\Omega^\Phi(v)$ is convex (even if $\Omega(p)$ is nonconvex).
3. **Order reversing:** if $\Omega(p) \leq \Lambda(p)$ for all $p \in \mathcal{C}$, then $\Omega^\Phi(v) \geq \Lambda^\Phi(v)$ for all $v \in \mathcal{V}$.
4. **Continuity:** Ω^Φ shares the same continuity modulus as Φ .
5. **Gradient (envelope theorem):** Under mild assumptions (see proof), we have
$$\nabla \Omega^\Phi(v) = \nabla_1 \Phi(v, p_\Omega^\Phi(v)),$$
 where ∇_1 denotes the gradient in the first argument.
6. **Smoothness:** If \mathcal{C} is a compact convex set, $\Phi(v, p)$ is β -smooth in (v, p) , concave in p and $\Omega(p)$ is γ -strongly convex in p , then $\Omega^\Phi(v)$ is $(\beta + \beta^2/\gamma)$ -smooth and $p_\Omega^\Phi(v)$ is β/γ -Lipschitz.

Generalized Fenchel-Young losses

(Blondel et al, 2022)

$$L_{\Omega}^{\Phi}(v, y) := \Omega^{\Phi}(v) + \Omega(y) - \Phi(v, y)$$



GFY loss properties

Non-negativity

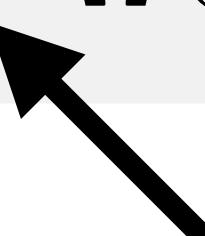
$$L_{\Omega}^{\Phi}(v, y) \geq 0$$

Zero loss

$$L_{\Omega}^{\Phi}(v, y) = 0 \Leftrightarrow p_{\Omega}^{\Phi}(v) = y$$

Gradient

$$\begin{aligned}\nabla_1 L_{\Omega}^{\Phi}(v, p) &= \nabla \Omega^{\Phi}(v) - \nabla_1 \Phi(v, y) \\ &= \nabla_1 \Phi(v, p_{\Omega}^{\Phi}(v)) - \nabla_1 \Phi(v, y)\end{aligned}$$



envelope theorems

Bounds

Lower bound

$\Phi(v, p) - \Omega(p)$ γ -strongly concave in p



$$\frac{\gamma}{2} \|p - p_{\Omega}^{\Phi}(v)\|^2 \leq L_{\Omega}^{\Phi}(v, p)$$

Upper bound

$\Phi(v, p)$ concave in p



$$L_{\Omega}^{\Phi}(v, p) \leq L_{\Omega}(\nabla_2 \Phi(v, p), p)$$

Multilabel classification experiments

Energy comparison (test accuracy in %)

Energy	yeast	scene	mediamill	birds	emotions	cal500
Unary (linear)	79.76	89.14	96.84	86.47	78.22	85.67
Unary (rectifier network)	80.03	91.35	96.91	91.74	79.79	86.25
Pairwise	80.19	91.58	96.95	91.55	80.56	85.73
SPEN	79.99	91.24	96.68	91.41	79.35	86.25
Input-concave SPEN	80.00	90.64	96.95	91.77	79.73	86.35

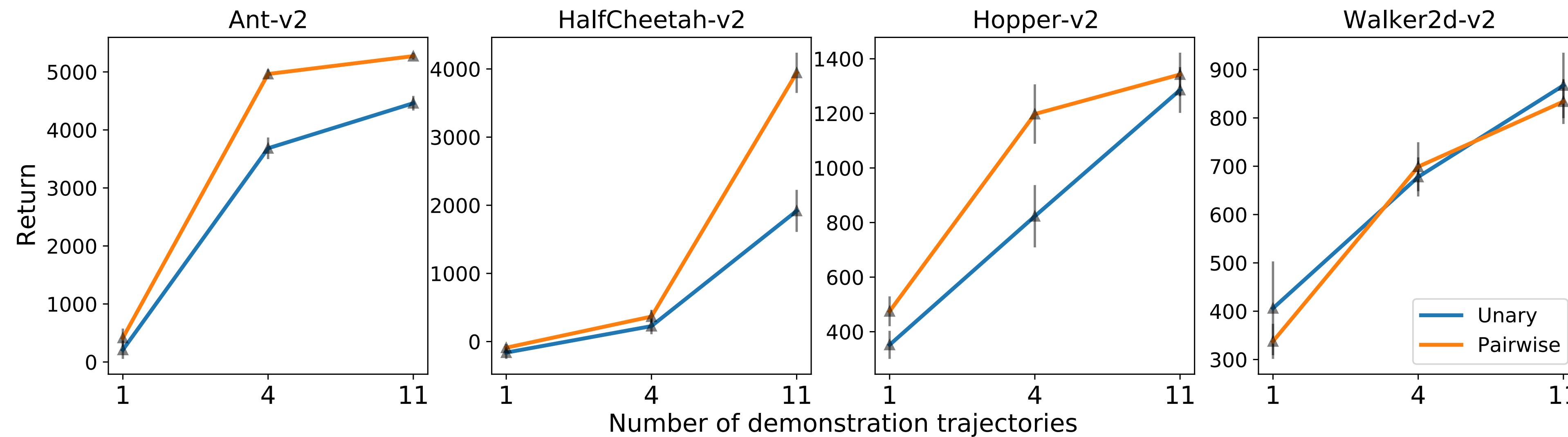
\mathcal{X} : features

$\mathcal{Y} = \{0, 1\}^k$: multiple labels

Loss comparison (test accuracy in %)

	yeast	scene	mediamill	birds	emotions	cal500
Generalized FY loss	80.19	91.58	96.95	91.55	80.56	85.73
Energy loss	42.35	33.02	40.92	14.29	55.50	39.27
Cross-entropy loss	79.00	90.78	96.77	91.56	78.08	85.89
Generalized perceptron loss	68.36	89.33	93.24	88.92	66.34	80.11

Imitation learning experiments



\mathcal{X} : current observations / state

\mathcal{Y} : angle of the arm joints in $[0, 1]^k$

Thank you