Learning Energy Networks with Generalized Fenchel-Young Losses


Mathieu Blondel


Felipe Llinares


Robert Dadashi

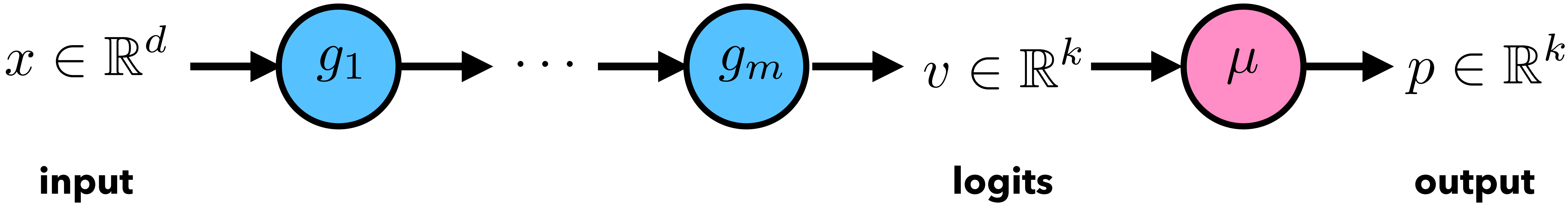

Léonard Hussenot


Matthieu Geist

SMAI-MODE workshop
Limoges, June 1st, 2022

Google Research

Neural networks

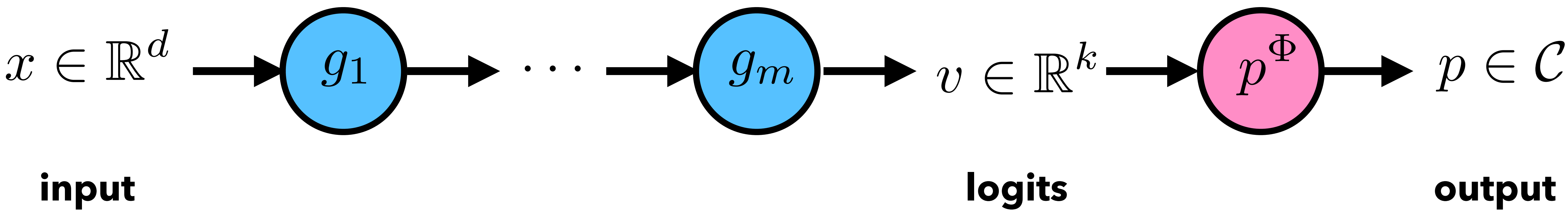


$$v = g(x) = (g_m \circ \cdots \circ g_1)(x) \in \mathbb{R}^k$$ **feature layers**

$$p = \mu(v) \in \mathbb{R}^k$$ **output layer (closed form)**

Energy networks

(LeCun, 2006)



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

output maximizing an energy Φ

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

Existing losses for energy networks

(LeCun, 2006)

Energy loss

$$(v, y) \mapsto -\Phi(v, y)$$

Can only work well if the energy is a negated loss

Composite loss

$$(v, y) \mapsto L(p^{\Phi}(v), y) \qquad L \colon \mathcal{C} \times \mathcal{Y} \to \mathbb{R}_+$$

Costly-to-compute gradients in v

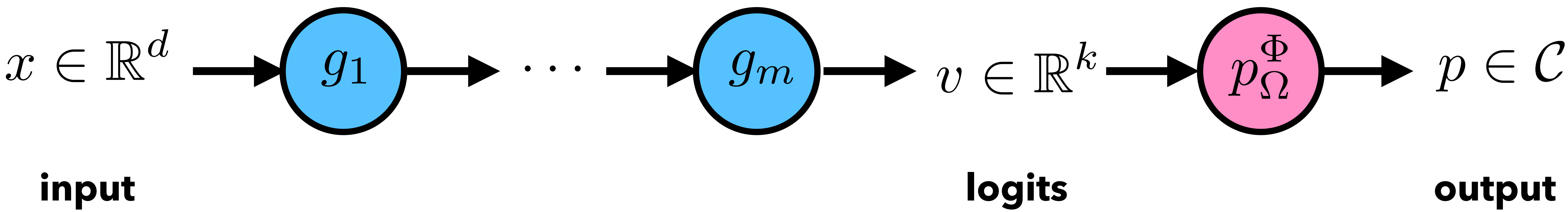
Generalized perceptron loss

$$(v, y) \mapsto \max_{p \in \mathcal{C}} \Phi(v, p) - \Phi(v, y)$$

Lack strong theoretical guarantees

Regularized energy networks

(Blondel et al, 2022)



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

output maximizing a regularized energy

$$\Phi \colon \mathcal{V} \times \mathcal{C} \to \mathbb{R} \qquad \Omega \colon \mathcal{C} \to \mathbb{R}$$

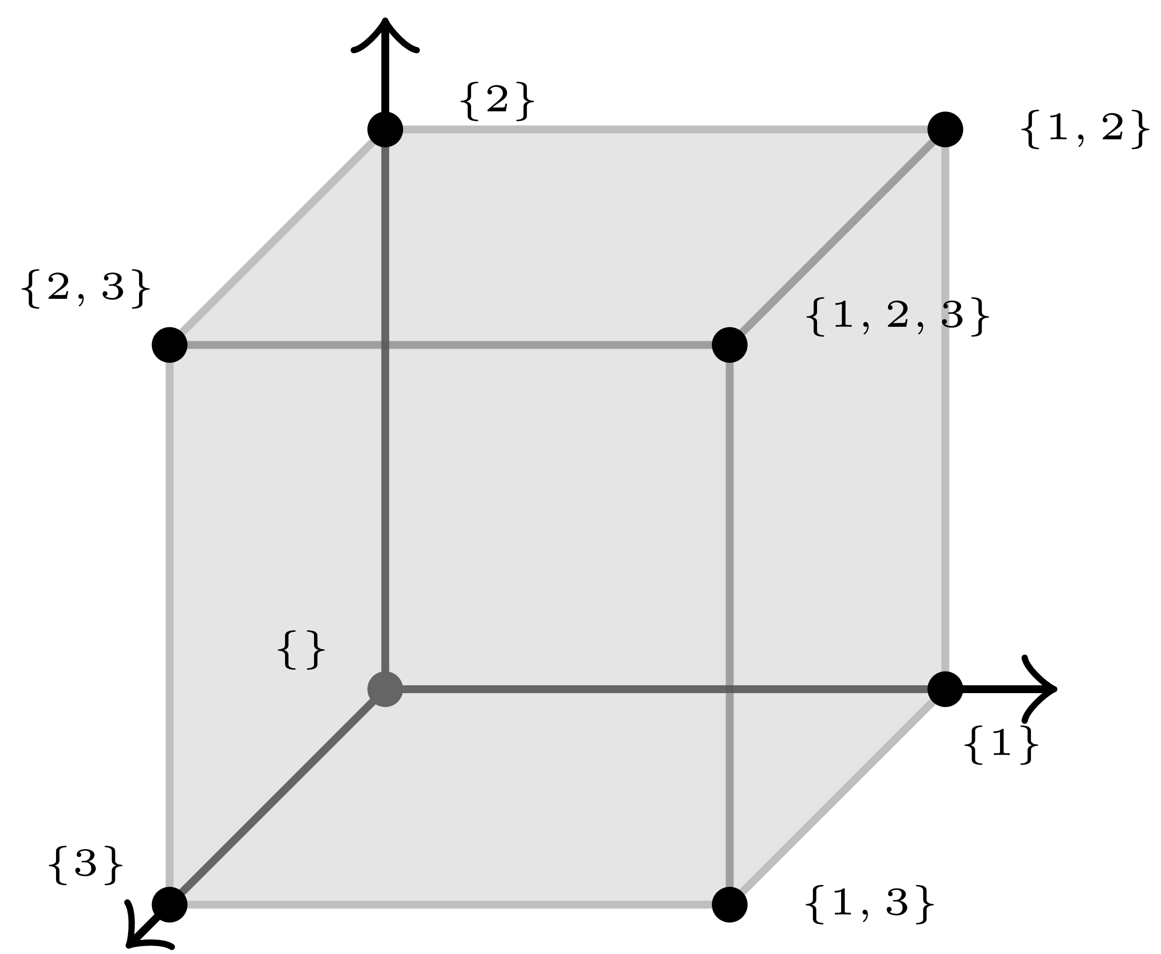
We propose a new loss construction for learning such networks

Example: linear-quadratic energy

(Blondel et al, 2022)

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

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

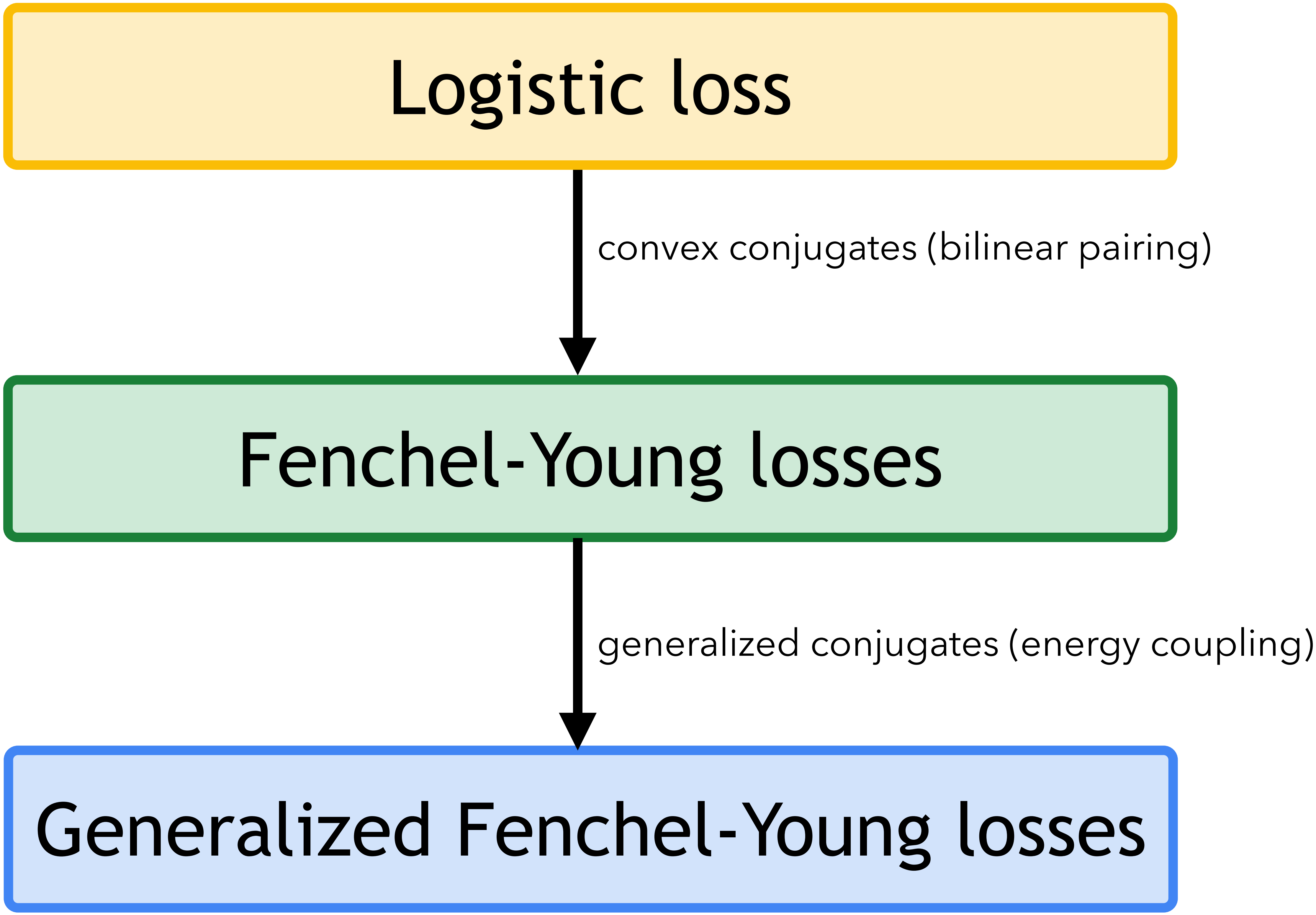


Regularized energy networks

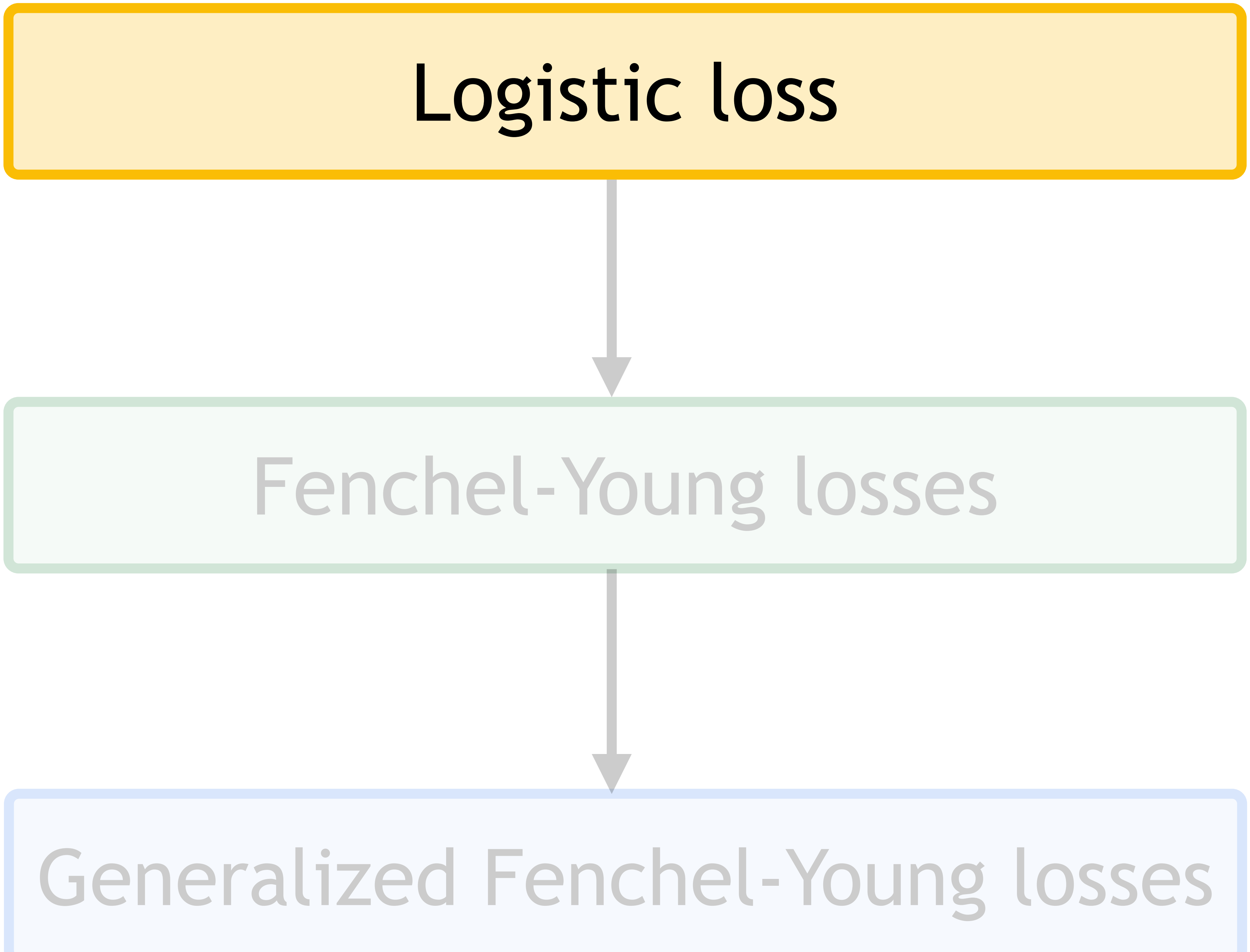
(Blondel et al, 2022)

	$\Phi(v,p)$	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	$\langle \mathrm{relu}(v), Up\rangle$	convex	linear
Maxout network	$p \cdot \max(v)$	convex	linear
LSE network	$p \cdot \mathrm{LSE}^\gamma(v)$	convex	linear
ICNN	$-\mathrm{ICNN}(v,p)$	nonconvex	concave
Probabilistic	$\sum_{y\in\mathcal{Y}} p(y)E(v,y)$	nonconvex	linear
Arbitrary	$\Phi(v,p)$	nonconvex	nonconcave

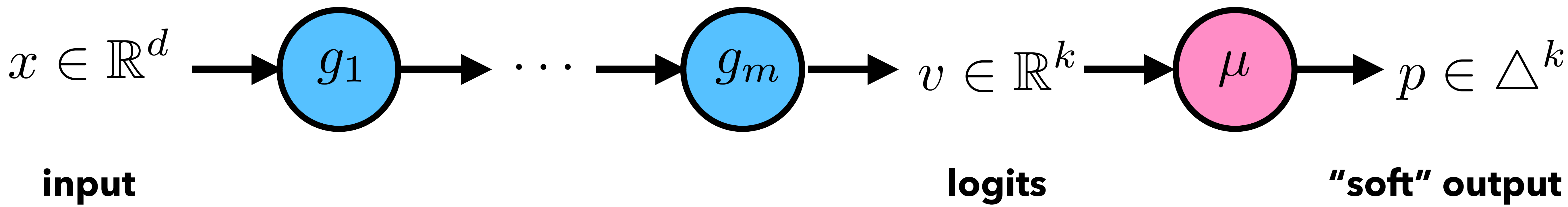
Remainder of this talk



Outline


Logistic loss
Fenchel-Young losses
Generalized Fenchel-Young losses

Neural networks with softmax output layer



$$v = g(x) = (g_m \circ \cdots \circ g_1)(x) \in \mathbb{R}^k$$

feature layers

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

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

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

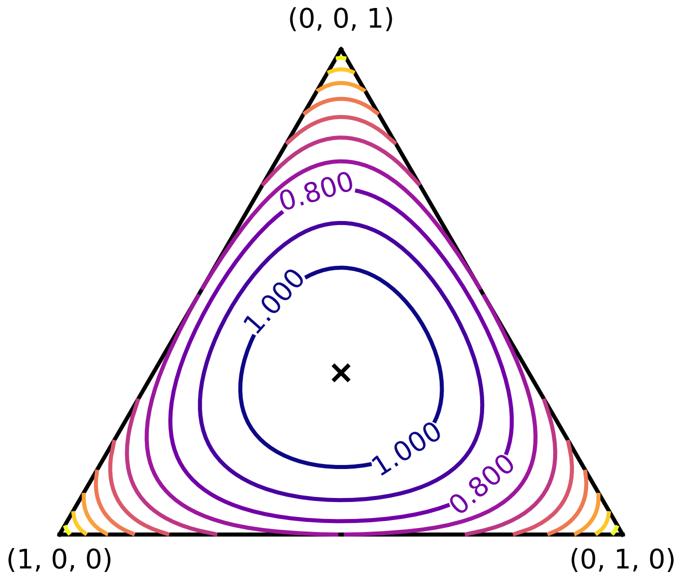
softmax $p = \mu(v) = \frac{\exp(v)}{\sum_{j=1}^{k} \exp(v_j)} \in \triangle^k$

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

Softmax as an argmax output layer

$$\mathrm{logsumexp}(v) = \log \sum_{j=1}^{k} \exp(v_j)$$

$$= \max_{p \in \triangle^k} \langle v, p \rangle - \langle p, \log p \rangle$$

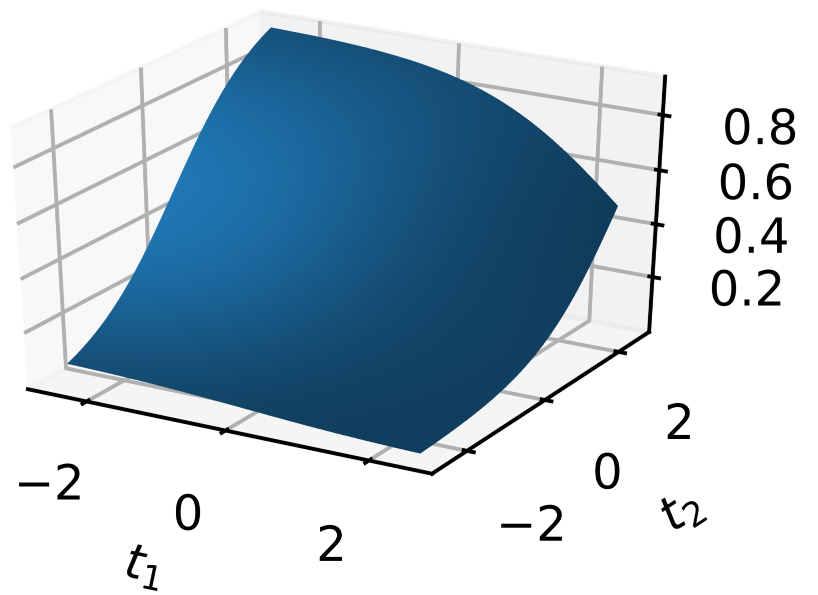


Contours of Shannon's entropy

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

$$= \operatorname{argmax}_{p \in \triangle^k} \langle v, p \rangle - \langle p, \log p \rangle$$

$$= \nabla \mathrm{logsumexp}(v)$$



Surface of the softmax

$\mu(t_1, t_2, 0)$

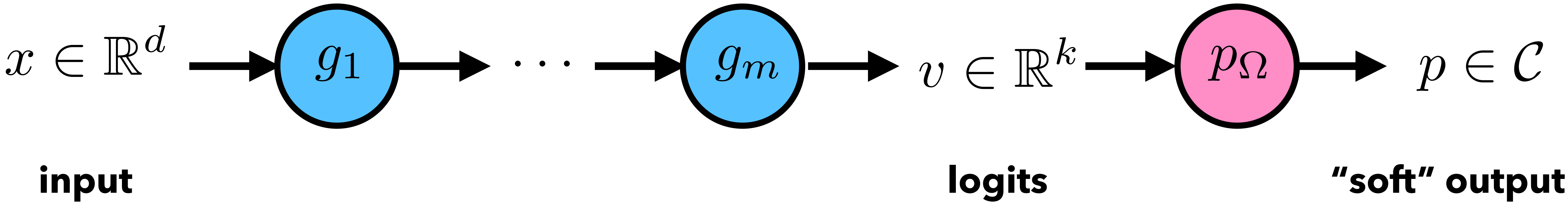
Outline

Logistic loss

Fenchel-Young losses

Generalized Fenchel-Young losses

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

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

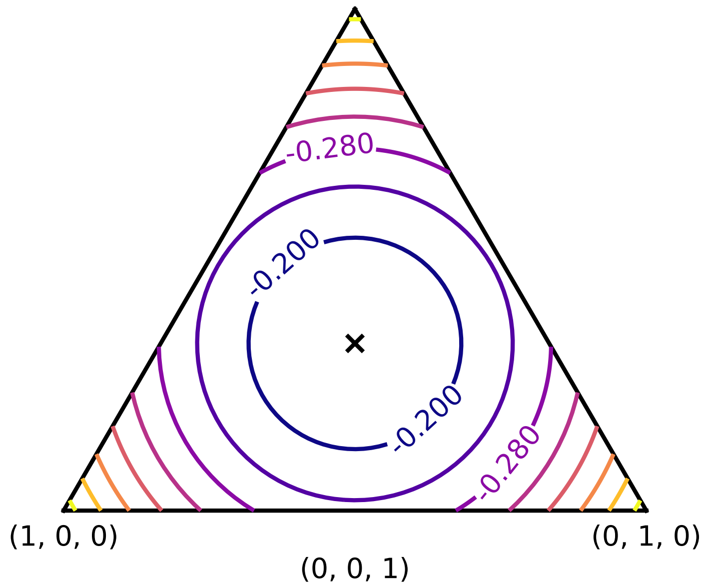
$$\Omega \colon \mathcal{C} \to \mathbb{R}$$

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

$f(v)$ is closed and convex

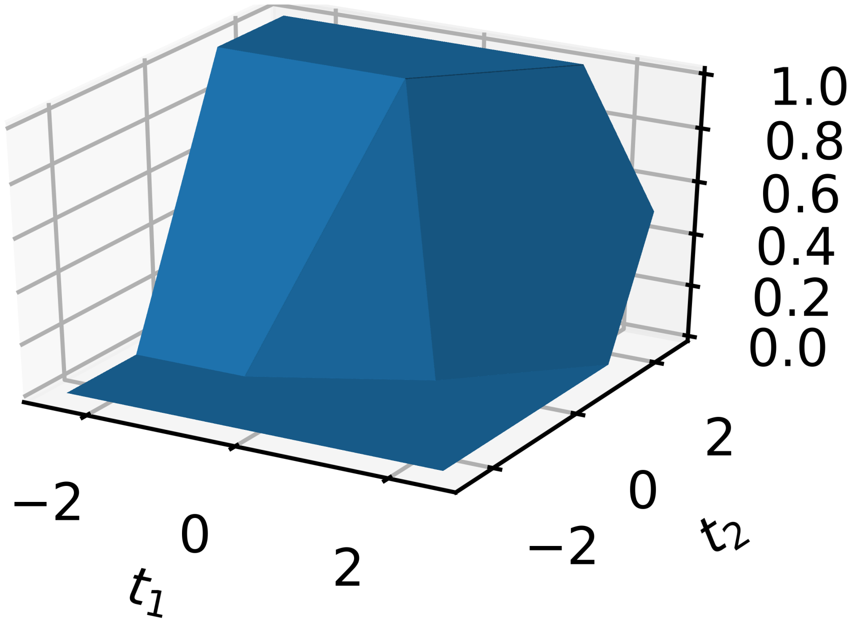
$$\Updownarrow$$

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



Contours of Gini's entropy

$\Omega(p) = 1/2 \langle p, 1 - p \rangle$



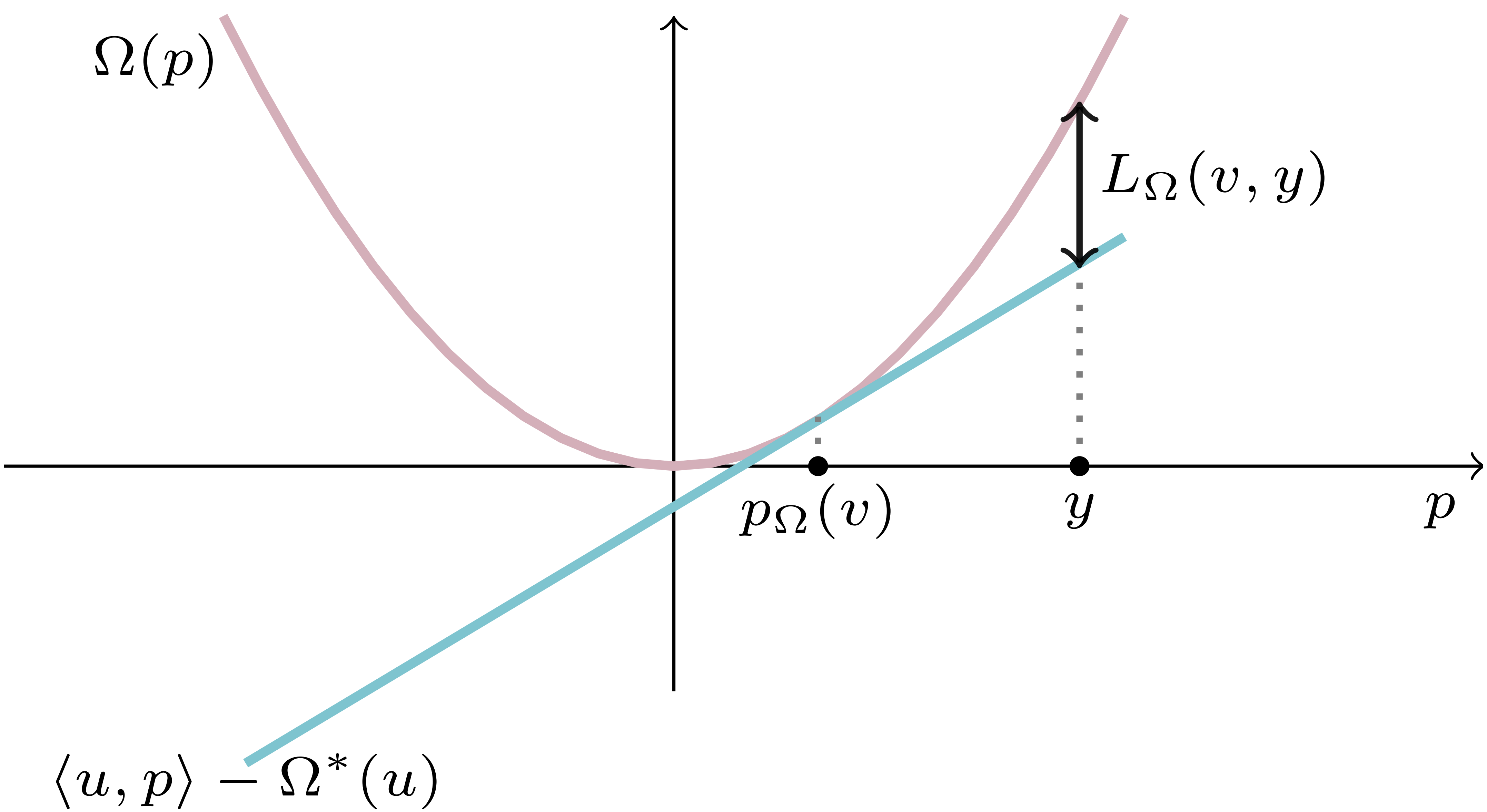
Surface of the sparsemax

$p_\Omega(t_1, t_2, 0)$

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

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

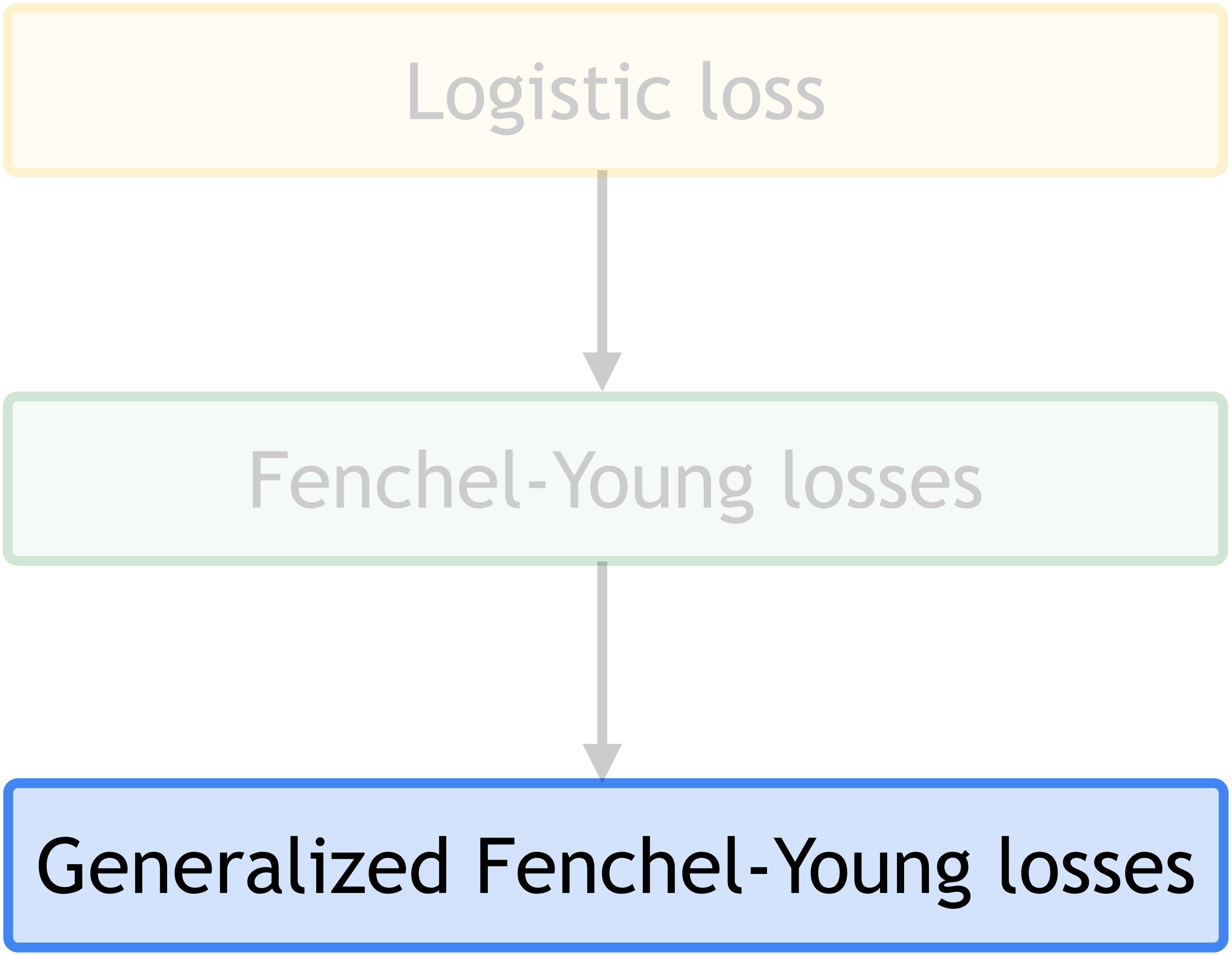
Zero loss

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

Gradient

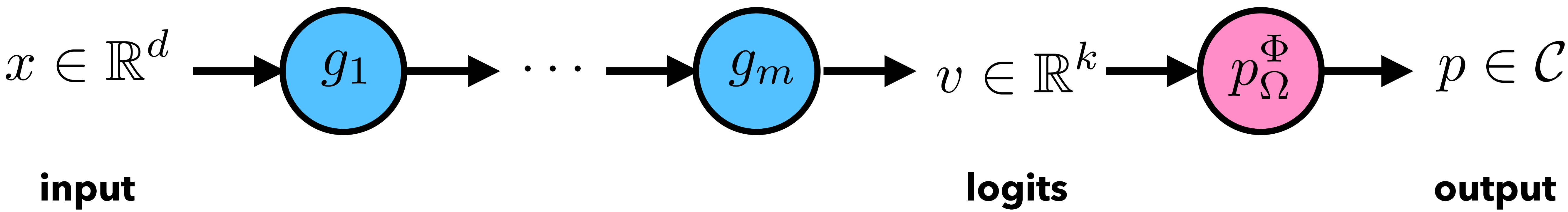
$$\begin{aligned} \nabla_1 L_\Omega(v, p) &= \nabla \Omega^*(v) - y \\ &= p_\Omega(v) - y \end{aligned}$$

Outline


Logistic loss
Fenchel-Young losses
Generalized Fenchel-Young losses

Regularized energy networks

(Blondel et al, 2022)



$$p = p_\Omega^\Phi(v) = \underset{p \in \mathcal{C}}{\text{argmax}} \; \Phi(v, p) - \Omega(p) \in \mathcal{C}$$

output maximizing a regularized energy

$$\Phi \colon \mathcal{V} \times \mathcal{C} \to \mathbb{R} \qquad \Omega \colon \mathcal{C} \to \mathbb{R}$$

Generalized conjugates

(Moreau, 1966)

$\Phi \colon \mathcal{V} \times \mathcal{C} \to \mathbb{R} \qquad \Omega \colon \mathcal{C} \to \mathbb{R}$

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

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

$$F(v) \text{ is } \Phi\text{-convex}$$

$$\Updownarrow$$

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

Generalized conjugate properties

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

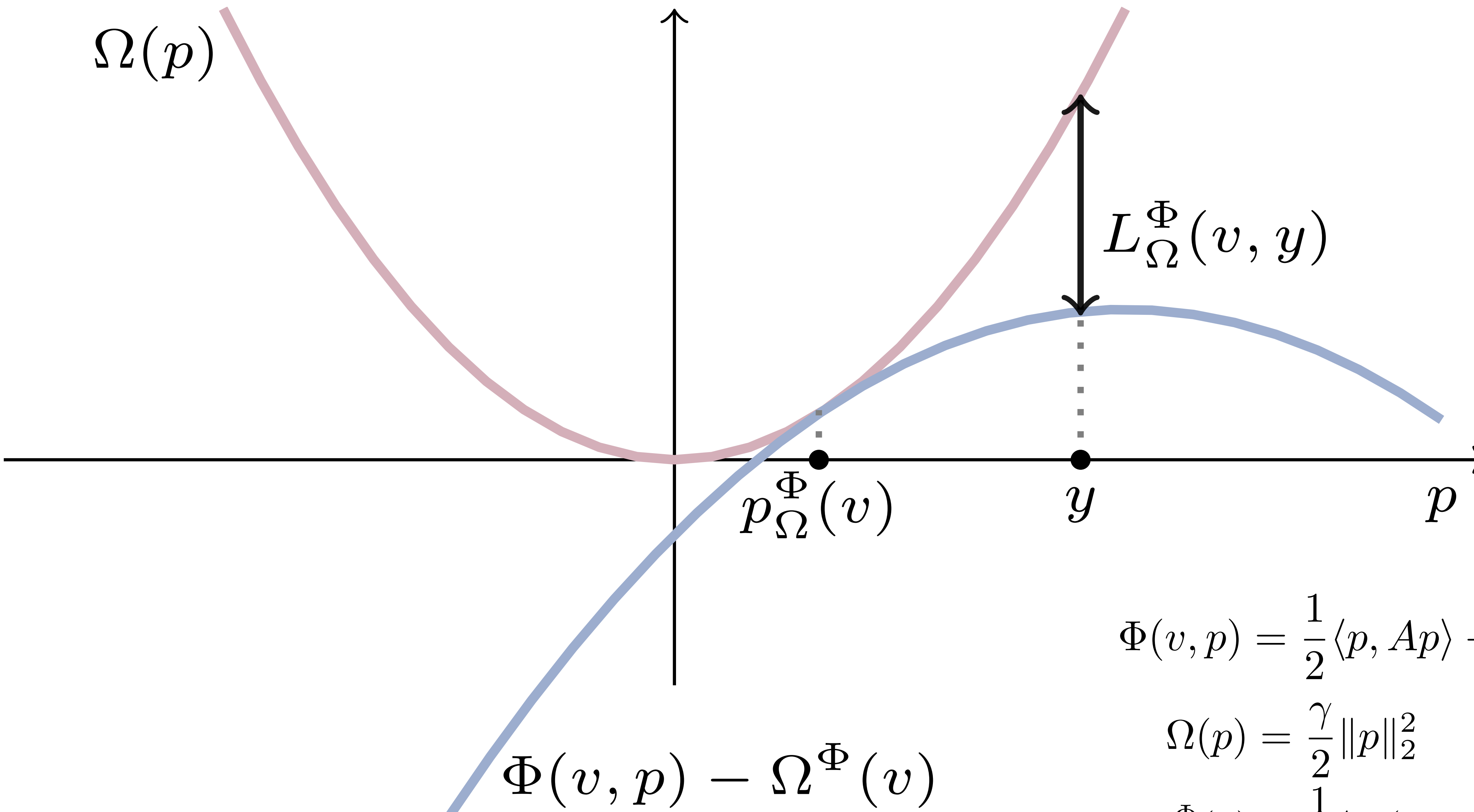
$$\Omega^{\Phi}(v) + \Omega(p) - \Phi(v, p) \geq 0.$$

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)

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



$$\Phi(v, p) = \frac{1}{2}\langle p, Ap\rangle + \langle b, p\rangle$$

$$\Omega(p) = \frac{\gamma}{2}\|p\|_2^2$$

$$\Omega^\Phi(v) = \frac{1}{2}\langle b, (\gamma I - A)^{-1} b\rangle$$

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

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

Upper bound

$$\Phi(v,p) \text{ concave in } p$$
$$\Downarrow$$
$$L_\Omega^\Phi(v,p) \leq L_\Omega(\nabla_2\Phi(v,p),p)$$

Multilabel classification experiments

Energy comparison (test accuracy in %)

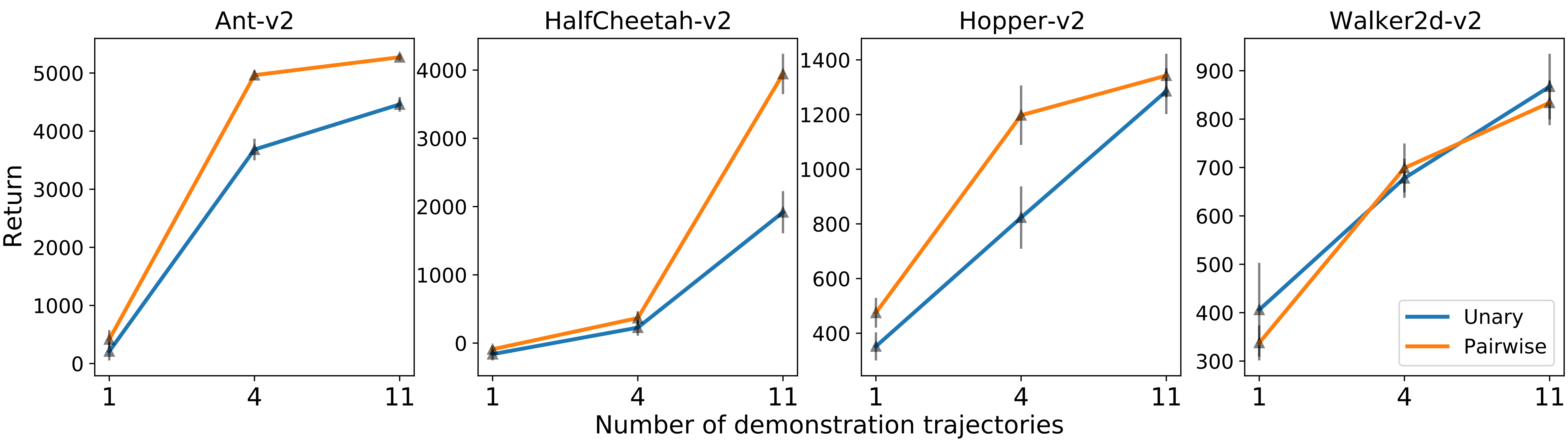
Energy	yeast	scene	mediamill	birds	emotions	cal500
Unary (linear)	79.76	89.14	96.84	86.47	78.22	85.67
Unary (rectifier network)	80.03	91.35	96.91	91.74	79.79	86.25
Pairwise	**80.19**	**91.58**	**96.95**	91.55	**80.56**	85.73
SPEN	79.99	91.24	96.68	91.41	79.35	86.25
Input-concave SPEN	80.00	90.64	**96.95**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



$$\mathcal{X}: \text{ current observations / state}$$

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

Conclusion

- We proposed a principled loss construction based on **general conjugates** for learning energy networks

- Preprint "Learning Energy Networks with Generalized Fenchel-Young losses" arXiv:2205.09589

  - More properties
  - Link with C-transforms
  - More experiments
  - Calibration guarantees in the surrogate loss setting
  - Generalized Bregman divergences

- Thank you for your attention!