Structured Output Support Vector Machines

Czech Technical University in Prague
V. Franc

March 29, 2022

- Margin-rescaling loss
- Structured Output Support Vector Machines

XEP33SML – Structured Model Learning, Summer 2022

Structured Output SVM

- Learning $h(x; \boldsymbol{w}) = \mathrm{Argmax}_{y \in \mathcal{Y}} \langle \boldsymbol{w}, \boldsymbol{\phi}(x, y) \rangle$ from examples $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ by ERM leads to

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w} \in \mathbb{R}^n} R_{\mathcal{T}^m}(\boldsymbol{w}) \quad \text{where} \quad R_{\mathcal{T}^m}(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell(y^i, h(x^i; \boldsymbol{w}))$$

Structured Output SVM

- Learning $h(x;\boldsymbol{w}) = \mathrm{Argmax}_{y\in\mathcal{Y}}\langle \boldsymbol{w}, \boldsymbol{\phi}(x,y)\rangle$ from examples $\mathcal{T}^m = \{(x^i,y^i)\in(\mathcal{X}\times\mathcal{Y}) \mid i=1,\ldots,m\}$ by ERM leads to

$$\boldsymbol{w}^* \in \mathop{\mathrm{Argmin}}_{\boldsymbol{w}\in\mathbb{R}^n} R_{\mathcal{T}^m}(\boldsymbol{w}) \quad \text{where} \quad R_{\mathcal{T}^m}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^m \ell(y^i, h(x^i;\boldsymbol{w}))$$

- The SO-SVM approximates the ERM by a convex problem

$$\boldsymbol{w}^* \in \mathop{\mathrm{Argmin}}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R^\psi(\boldsymbol{w})\right) \quad \text{where} \quad R^\psi(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^m \psi(x^i, y^i, \boldsymbol{w})$$

Structured Output SVM

- Learning $h(x;\boldsymbol{w}) = \operatorname{Argmax}_{y\in\mathcal{Y}}\langle \boldsymbol{w}, \boldsymbol{\phi}(x,y)\rangle$ from examples $\mathcal{T}^m = \{(x^i,y^i)\in(\mathcal{X}\times\mathcal{Y}) \mid i=1,\ldots,m\}$ by ERM leads to

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} R_{\mathcal{T}^m}(\boldsymbol{w}) \quad \text{where} \quad R_{\mathcal{T}^m}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^m \ell(y^i, h(x^i;\boldsymbol{w}))$$

- The SO-SVM approximates the ERM by a convex problem

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R^\psi(\boldsymbol{w})\right) \quad \text{where} \quad R^\psi(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^m \psi(x^i, y^i, \boldsymbol{w})$$

- The surrogate loss $\psi\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R}^n \to \mathbb{R}$ is an upper bound:

$$\ell(y, h(x;\boldsymbol{w})) \le \psi(x,y,\boldsymbol{w})\,, \quad \forall (x,y,\boldsymbol{w}) \in (\mathcal{X}\times\mathcal{Y}\times\mathbb{R}^n)$$

which is convex in $\boldsymbol{w}$ for any (x,y).

Margin rescaling loss

- We require the score of the correct label y^i to be higher than the score of any incorrect label y by margin proportional to the loss $\ell(y^i, y)$:

$$\langle \boldsymbol{w}, \phi(x^i, y^i) \rangle \geq \langle \boldsymbol{w}, \phi(x^i, y) \rangle + \ell(y^i, y)\,, \qquad \forall y \in \mathcal{Y} \setminus \{y^i\}$$

Margin rescaling loss

- We require the score of the correct label y^i to be higher than the score of any incorrect label y by margin proportional to the loss $\ell(y^i, y)$:

$$\langle \boldsymbol{w}, \phi(x^i, y^i) \rangle \geq \langle \boldsymbol{w}, \phi(x^i, y) \rangle + \ell(y^i, y) \,, \qquad \forall y \in \mathcal{Y} \setminus \{y^i\}$$

- Example: Sequencial OCR, Hamming distance $\ell(\boldsymbol{y}, \boldsymbol{y}') = \sum_{i=1}^{L} [\![y_i \neq y_i']\!]$

$$\begin{aligned}
\Big\langle \boldsymbol{\phi}\Big(\text{JOHN}, JOHN\Big), \boldsymbol{w} \Big\rangle &\geq \Big\langle \boldsymbol{\phi}\Big(\text{JOHN}, AAAA\Big), \boldsymbol{w} \Big\rangle + 4 \\
\Big\langle \boldsymbol{\phi}\Big(\text{JOHN}, JOHN\Big), \boldsymbol{w} \Big\rangle &\geq \Big\langle \boldsymbol{\phi}\Big(\text{JOHN}, JAAA\Big), \boldsymbol{w} \Big\rangle + 3 \\
\Big\langle \boldsymbol{\phi}\Big(\text{JOHN}, JOHN\Big), \boldsymbol{w} \Big\rangle &\geq \Big\langle \boldsymbol{\phi}\Big(\text{JOHN}, JOAA\Big), \boldsymbol{w} \Big\rangle + 2 \\
\Big\langle \boldsymbol{\phi}\Big(\text{JOHN}, JOHN\Big), \boldsymbol{w} \Big\rangle &\geq \Big\langle \boldsymbol{\phi}\Big(\text{JOHN}, JOHA\Big), \boldsymbol{w} \Big\rangle + 1
\end{aligned}$$

$$\vdots$$

Margin rescaling loss

- We require the score of the correct label y^i to be higher than the score of any incorrect label y by margin proportional to the loss $\ell(y^i, y)$:

$$\langle \boldsymbol{w}, \phi(x^i, y^i) \rangle \geq \langle \boldsymbol{w}, \phi(x^i, y) \rangle + \ell(y^i, y)\,, \qquad \forall y \in \mathcal{Y} \setminus \{y^i\}$$

- The margin rescaling loss

$$\psi(x^i, y^i, \boldsymbol{w}) = \max\left\{0, \max_{y \in \mathcal{Y} \setminus \{y^i\}} \big(\ell(y^i, y) + \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y) \rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y^i) \rangle \big)\right\}$$

Margin rescaling loss

- We require the score of the correct label y^i to be higher than the score of any incorrect label y by margin proportional to the loss $\ell(y^i, y)$:

$$\langle \boldsymbol{w}, \phi(x^i, y^i) \rangle \geq \langle \boldsymbol{w}, \phi(x^i, y) \rangle + \ell(y^i, y)\,, \qquad \forall y \in \mathcal{Y} \setminus \{y^i\}$$

- The margin rescaling loss

$$\psi(x^i, y^i, \boldsymbol{w}) = \max\Big\{0, \max_{y \in \mathcal{Y} \setminus \{y^i\}} \big(\ell(y^i, y) + \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y) \rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y^i) \rangle\big)\Big\}$$

- Upper bound of the true loss:

$$y^i \neq \hat{y} = h(x^i; \boldsymbol{w}) = \operatorname*{Argmax}_{y \in \mathcal{Y}} \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y) \rangle$$

implies $\langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, \hat{y}) \rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y^i) \rangle \geq 0$ and hence

$$\psi(x^i, y^i, \boldsymbol{w}) \geq \ell(y^i, h(x^i, \boldsymbol{w}))\,, \qquad \forall \boldsymbol{w} \in \mathbb{R}^n$$

Margin-rescaling loss

- Using shortcuts $\ell_i(y) = \ell(y^i, y)$ and $\boldsymbol{\phi}_i(y) = \boldsymbol{\phi}(x^i, y) - \boldsymbol{\phi}(x^i, y^i)$ we can simplify the margin rescaling loss:

$$
\begin{aligned}
\psi(x^i, y^i, \boldsymbol{w}) &= \max\{0, \max_{y \in \mathcal{Y} \setminus \{y^i\}} \big(\ell(y^i, y) + \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y)\rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y^i)\rangle\big)\} \\
&= \max_{y \in \mathcal{Y}} \big(\ell(y^i, y) + \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y)\rangle - \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y^i)\rangle\big) \\
&= \max_{y \in \mathcal{Y}} \big(\ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle\big)
\end{aligned}
$$

- The margin-rescaling loss is a point-wise maximum over $|\mathcal{Y}|$ linear terms, hence, it is convex.

SO-SVM leads to a convex QP

- The SO-SVM with margin-rescaling loss:

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max_{y\in\mathcal{Y}} \{\ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle\}}_{R^{\psi}(\boldsymbol{w})} \right)$$

SO-SVM leads to a convex QP

- The SO-SVM with margin-rescaling loss:

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max_{y\in\mathcal{Y}}\{\ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle\}}_{R^{\psi}(\boldsymbol{w})} \right)$$

- By using slack variables it can be rewritten as a Quadratic Program:

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n, \boldsymbol{\xi}\in\mathbb{R}^m} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i \right)$$

subject to

$$\xi_i \geq \ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle \,, \quad \forall i \in \{1,\ldots,m\}, \forall y \in \mathcal{Y}$$

- Note that the QP has $m|\mathcal{Y}|$ linear constaints !

Approximation of loss function

Theorem 1. *Let* $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ *be a loss such that* $\ell(y, y') = 0 \iff y = y'$*, and* $h(x) = \operatorname{argmax}_{y \in \mathcal{Y}} f(x, y)$ *a classifier* $h\colon \mathcal{X} \to \mathcal{Y}$*. Then*

$$\ell(h(x), y) \le \max_{y' \in \mathcal{Y} \setminus \{y\}} \psi\Big(f(x, y) - f(x, y'), \ell(y, y')\Big), \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

where $\psi\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *is a function such that* $\psi(t, u) \ge u\,[\![t \le 0]\!]$.

Approximation of loss function

Theorem 1. *Let* $\ell\colon \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}_+$ *be a loss such that* $\ell(y,y')=0 \iff y=y'$, *and* $h(x)=\operatorname{argmax}_{y\in\mathcal{Y}} f(x,y)$ *a classifier* $h\colon \mathcal{X}\to\mathcal{Y}$. *Then*

$$\ell(h(x),y) \le \max_{y'\in\mathcal{Y}\setminus\{y\}} \psi\Big(f(x,y)-f(x,y'),\ell(y,y')\Big)\,, \qquad \forall x\in\mathcal{X}, y\in\mathcal{Y}$$

where $\psi\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ *is a function such that* $\psi(t,u)\ge u\,[\![t\le 0]\!]$.

PROOF:

$$\begin{aligned}
\ell(h(x),y) &= \ell(h(x),y)\,[\![f(x,y)-\max_{y'\neq y} f(x,y')\le 0]\!] \\
&\le \psi\Big(f(x,y)-\max_{y'\neq y} f(x,y')\ ,\ \ell(h(x),y)\Big) \\
&= \psi\Big(f(x,y)-f(x,h(x))\ ,\ \ell(h(x),y)\Big) \\
&\le \max_{y'\neq y} \psi\Big(f(x,y)-f(x,y')\ ,\ \ell(y',y)\Big)
\end{aligned}$$

■

Approximation of loss function

- **Margin re-scaling loss:** $\psi(t,u) = \max\{0, u - t\}$

$$\ell(h(x),y) \leq \max_{y' \neq y} \max\left\{0, \ell(y',y) - f(x,y) + f(x,y')\right\}$$

$$= \max_{y'} \Big(\ell(y',y) - f(x,y) + f(x,y')\Big)$$

Approximation of loss function

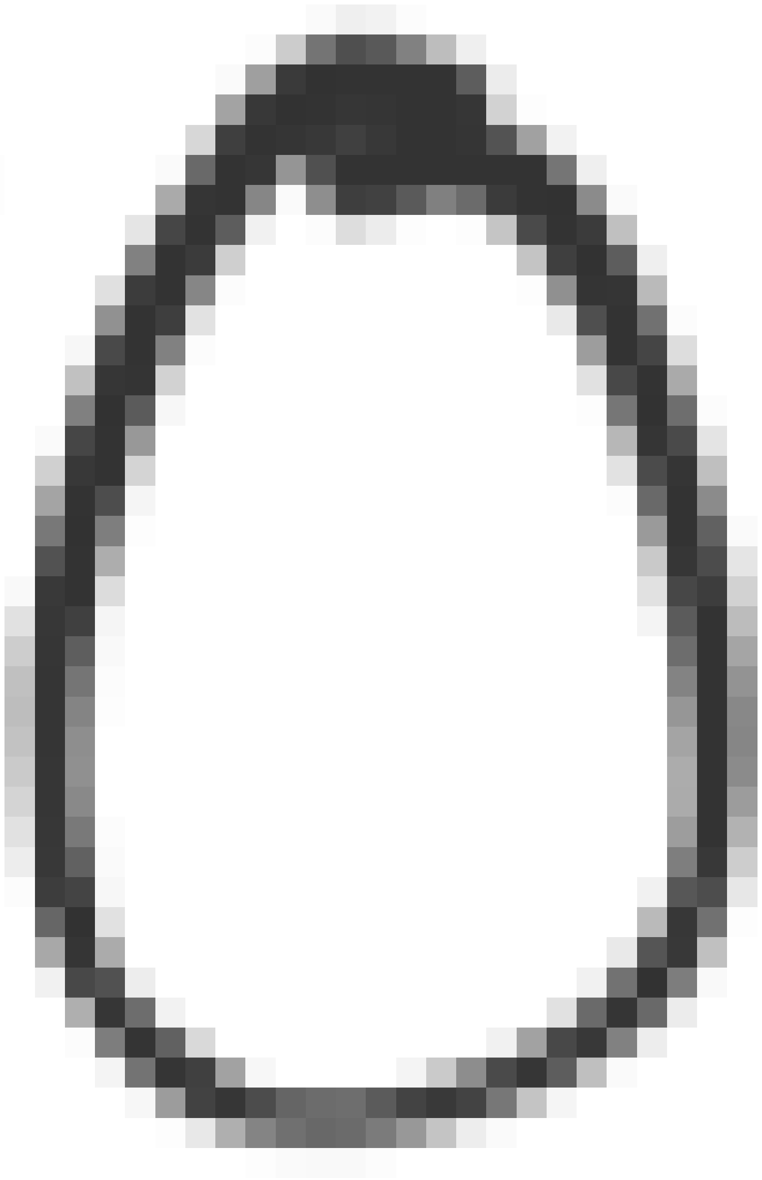
- **Margin re-scaling loss:** $\psi(t,u) = \max\{0, u - t\}$

$$
\begin{aligned}
\ell(h(x), y) &\leq \max_{y' \neq y} \max \Big\{ 0, \ell(y', y) - f(x,y) + f(x,y') \Big\} \\
&= \max_{y'} \Big(\ell(y', y) - f(x,y) + f(x,y') \Big)
\end{aligned}
$$

- **Slack re-scaling loss:** $\psi(t,u) = \max\{0, u\,(1 - t)\}$

$$
\begin{aligned}
\ell(h(x), y) &\leq \max_{y' \neq y} \max \Big\{ 0, \ell(y', y) \left(1 - f(x,y) + f(x,y')\right) \Big\} \\
&= \max_{y'} \ell(y', y) \Big(1 - f(x,y) + f(x,y') \Big)
\end{aligned}
$$

J



H

N

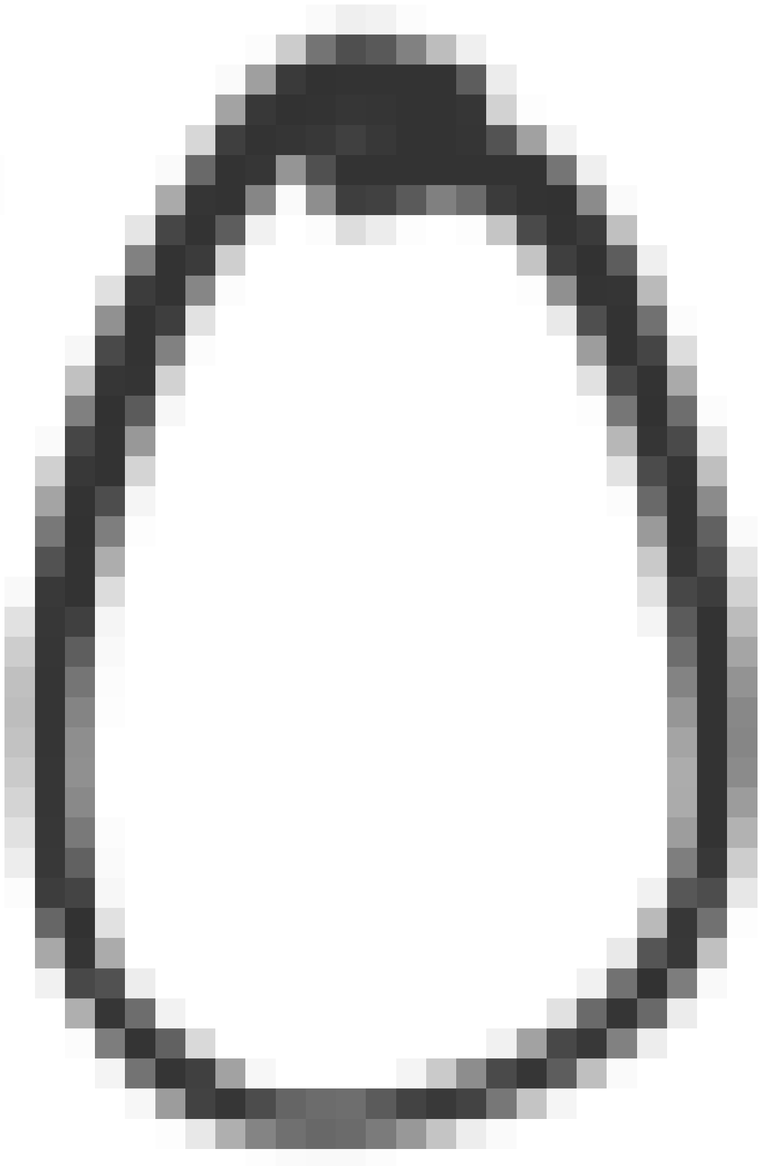
J

0

H

N

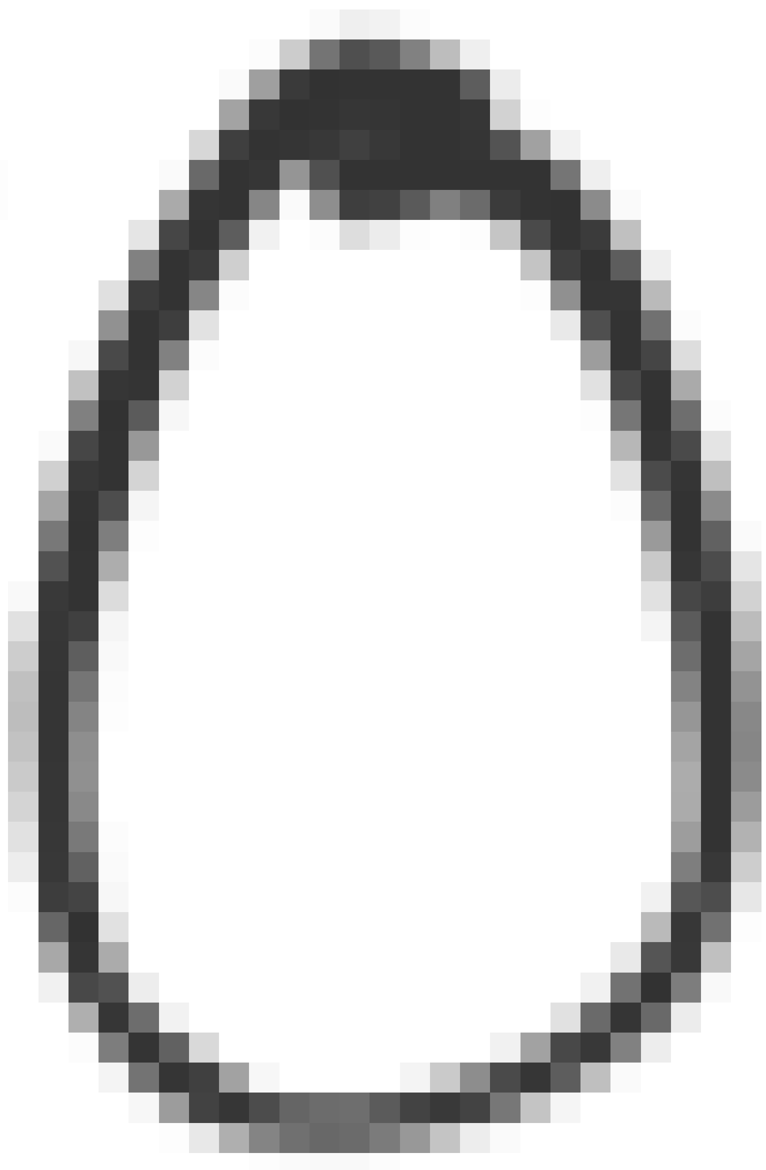
J



H

N

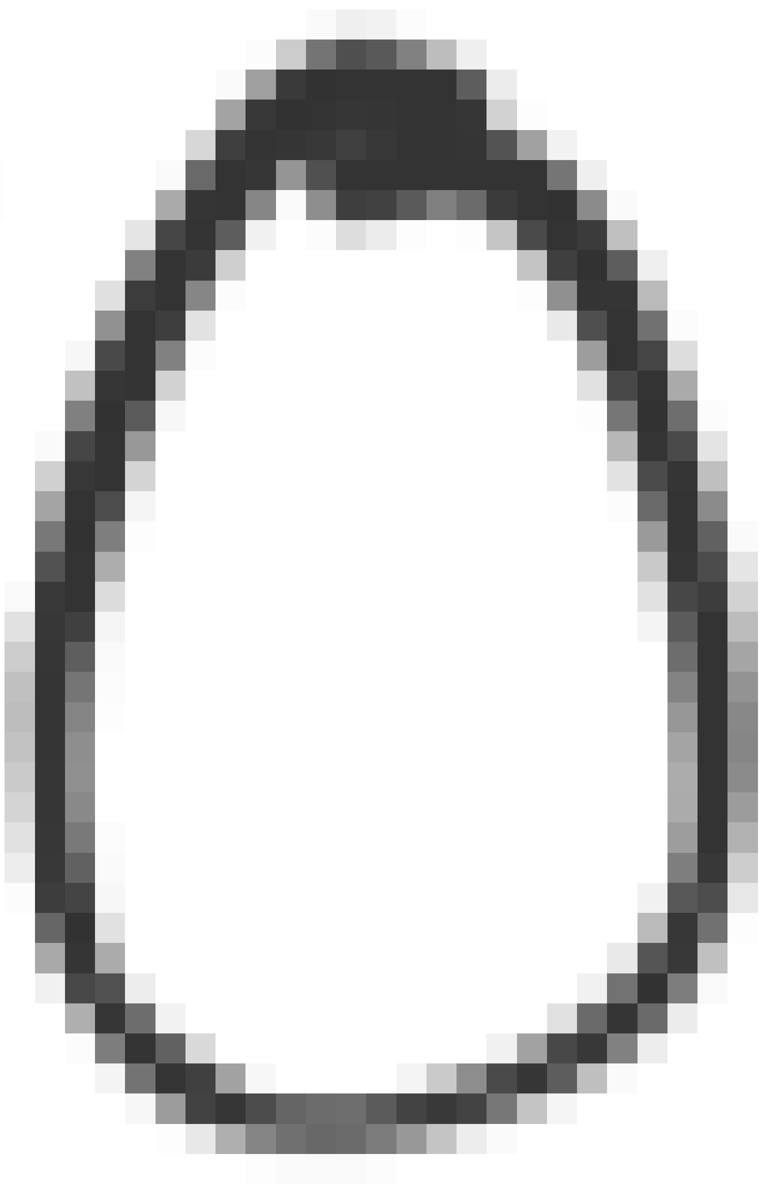
J



H

N

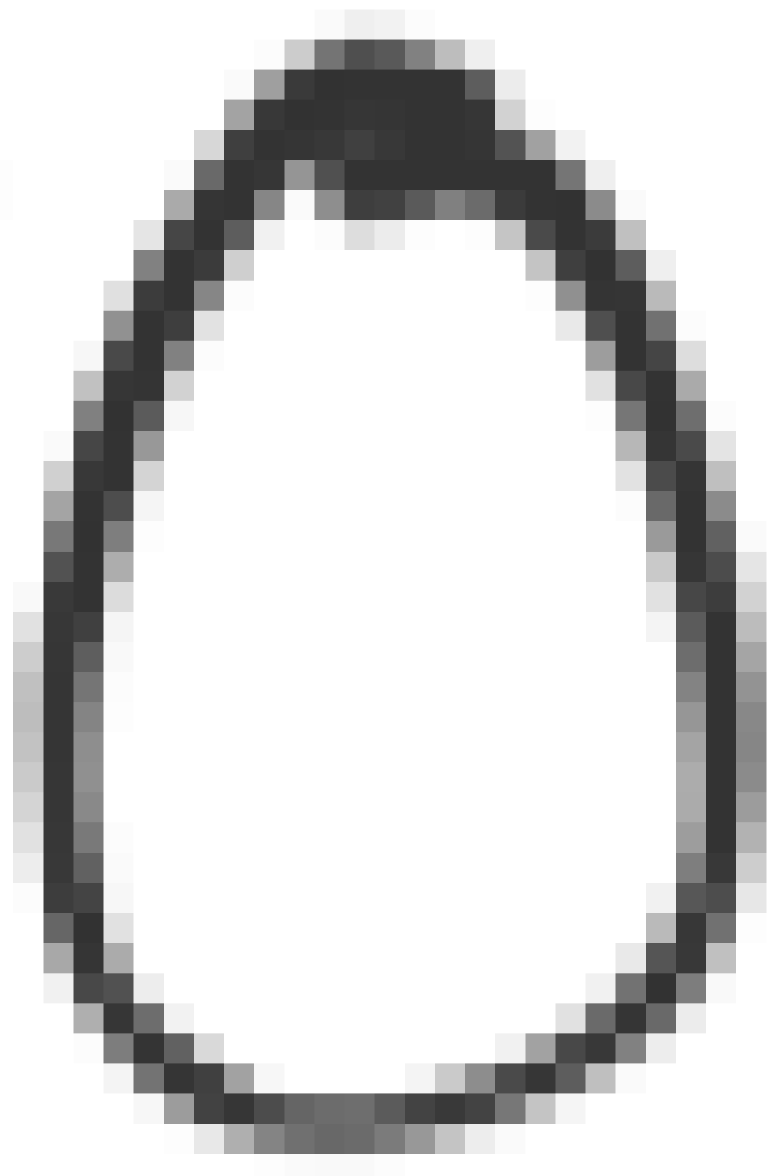
J



H

N

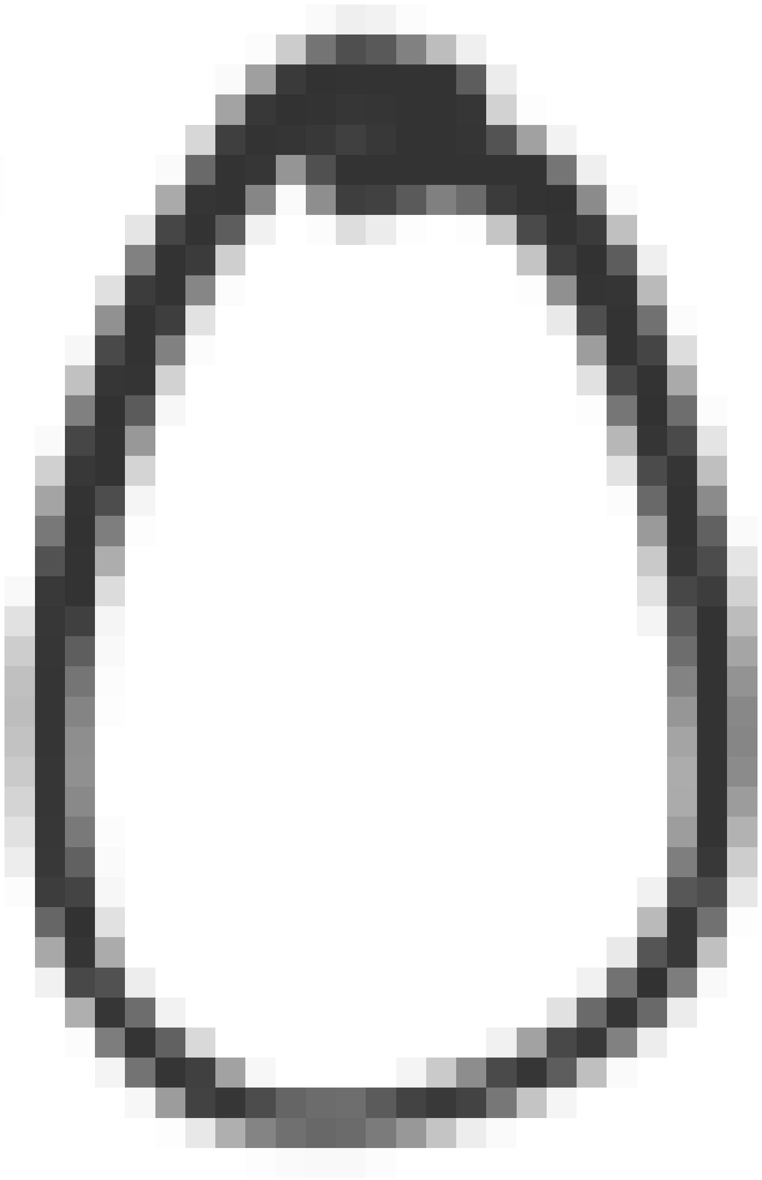
Ј



H

N

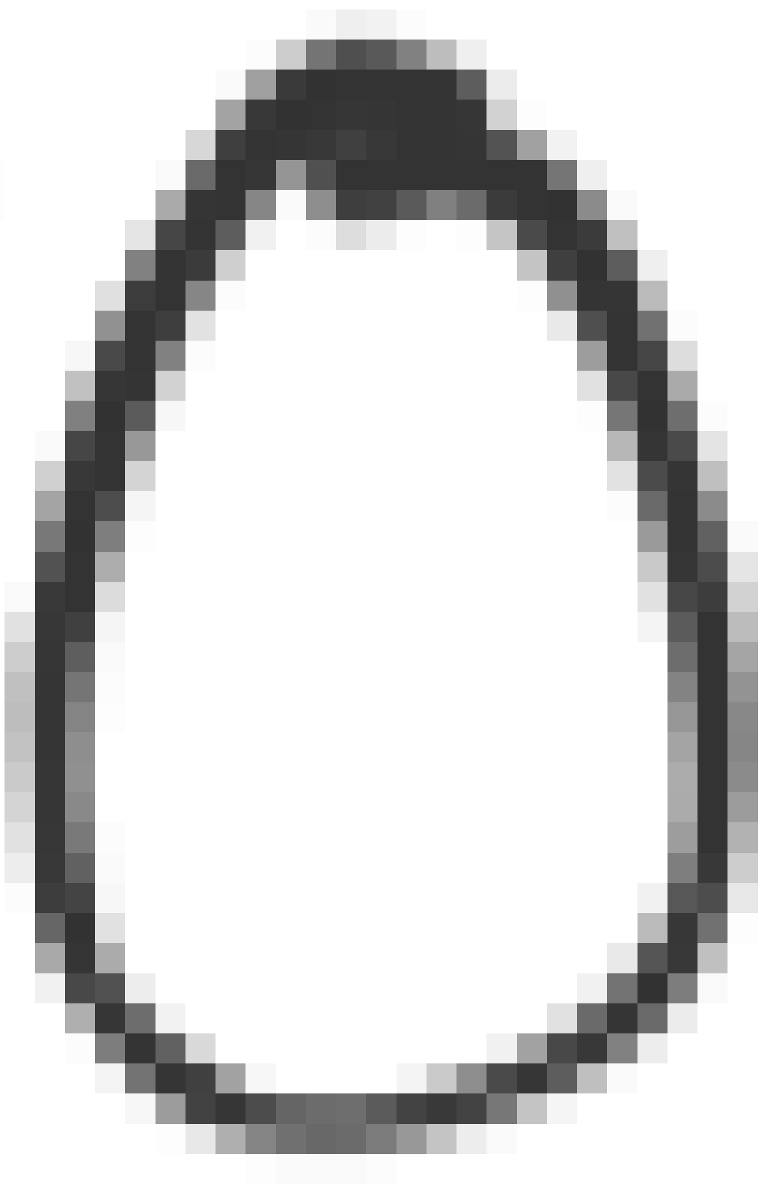
J



H

N

J



H

N