

Structured Output Support Vector Machines

Czech Technical University in Prague
V. Franc

February 24, 2020

- ◆ Margin-rescaling loss
- ◆ Structured Output Support Vector Machines

XEP33SML – Structured Model Learning, Summer 2020

Structured Output SVM

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

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

Structured Output SVM

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

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

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

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

Structured Output SVM

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

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

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

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

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

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

which is convex in \mathbf{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 $\ell(y^i, y)$:

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

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 $\ell(y^i, y)$:

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

- ◆ Example: Sequential OCR, Hamming distance $\ell(\mathbf{y}, \mathbf{y}') = \sum_{i=1}^L \mathbb{1}[y_i \neq y'_i]$

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

⋮

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 $\ell(y^i, y)$:

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

- ◆ The margin rescaling loss

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

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 $\ell(y^i, y)$:

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

- ◆ The margin rescaling loss

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

- ◆ Upper bound of the true loss:

$$y^i \neq \hat{y} = h(x^i; \mathbf{w}) = \underset{y \in \mathcal{Y}}{\text{Argmax}} \langle \mathbf{w}, \phi(x^i, y) \rangle$$

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

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

Margin-rescaling loss

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

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

- ◆ 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:

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

SO-SVM leads to a convex QP

- ◆ The SO-SVM with margin-rescaling loss:

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

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

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

subject to

$$\xi_i \geq \ell_i(y) + \langle \mathbf{w}, \phi_i(y) \rangle, \quad \forall i \in \{1, \dots, m\}, \forall y \in \mathcal{Y}$$

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

Approximation of loss function

Theorem 1. *Let $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ 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 $h: \mathcal{X} \rightarrow \mathcal{Y}$. Then*

$$\ell(h(x), y) \leq \max_{y' \in \mathcal{Y} \setminus \{y\}} \psi\left(f(x, y) - f(x, y'), \ell(y, y')\right), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

where $\psi: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a function such that $\psi(t, u) \geq u$ $\llbracket t \leq 0 \rrbracket$.

Approximation of loss function

Theorem 1. *Let $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ 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 $h: \mathcal{X} \rightarrow \mathcal{Y}$. Then*

$$\ell(h(x), y) \leq \max_{y' \in \mathcal{Y} \setminus \{y\}} \psi\left(f(x, y) - f(x, y'), \ell(y, y')\right), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

where $\psi: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a function such that $\psi(t, u) \geq u$ $\llbracket t \leq 0 \rrbracket$.

PROOF:

$$\begin{aligned} \ell(h(x), y) &= \ell(h(x), y) \llbracket f(x, y) - \max_{y' \neq y} f(x, y') \leq 0 \rrbracket \\ &\leq \psi\left(f(x, y) - \max_{y' \neq y} f(x, y'), \ell(h(x), y)\right) \\ &= \psi\left(f(x, y) - f(x, h(x)), \ell(h(x), y)\right) \\ &\leq \max_{y' \neq y} \psi\left(f(x, y) - f(x, y'), \ell(y', y)\right) \end{aligned}$$

Approximation of loss function

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

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

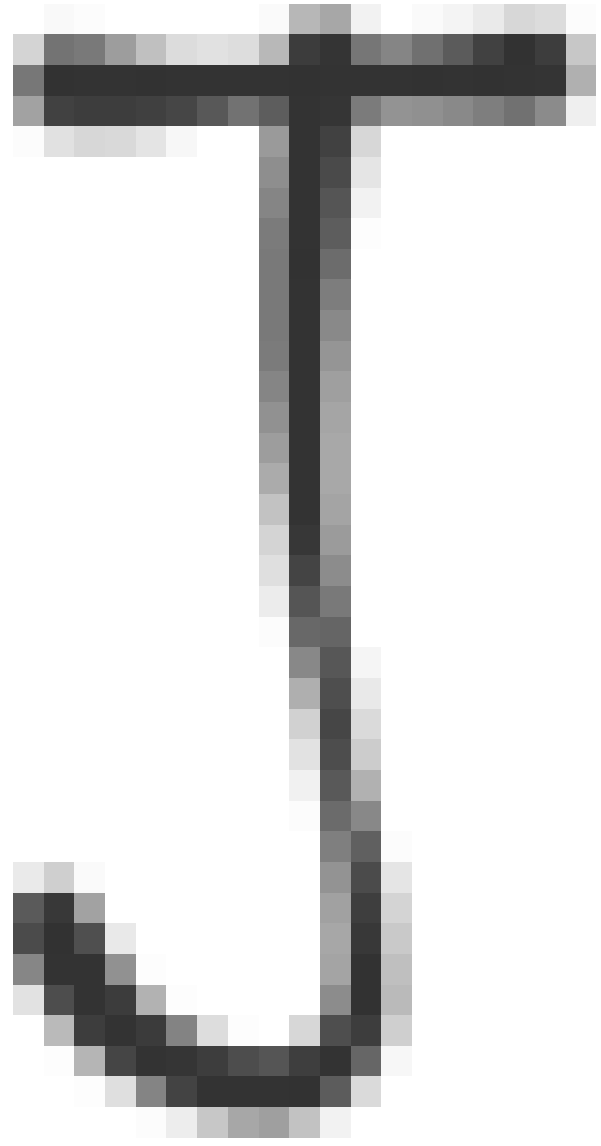
Approximation of loss function

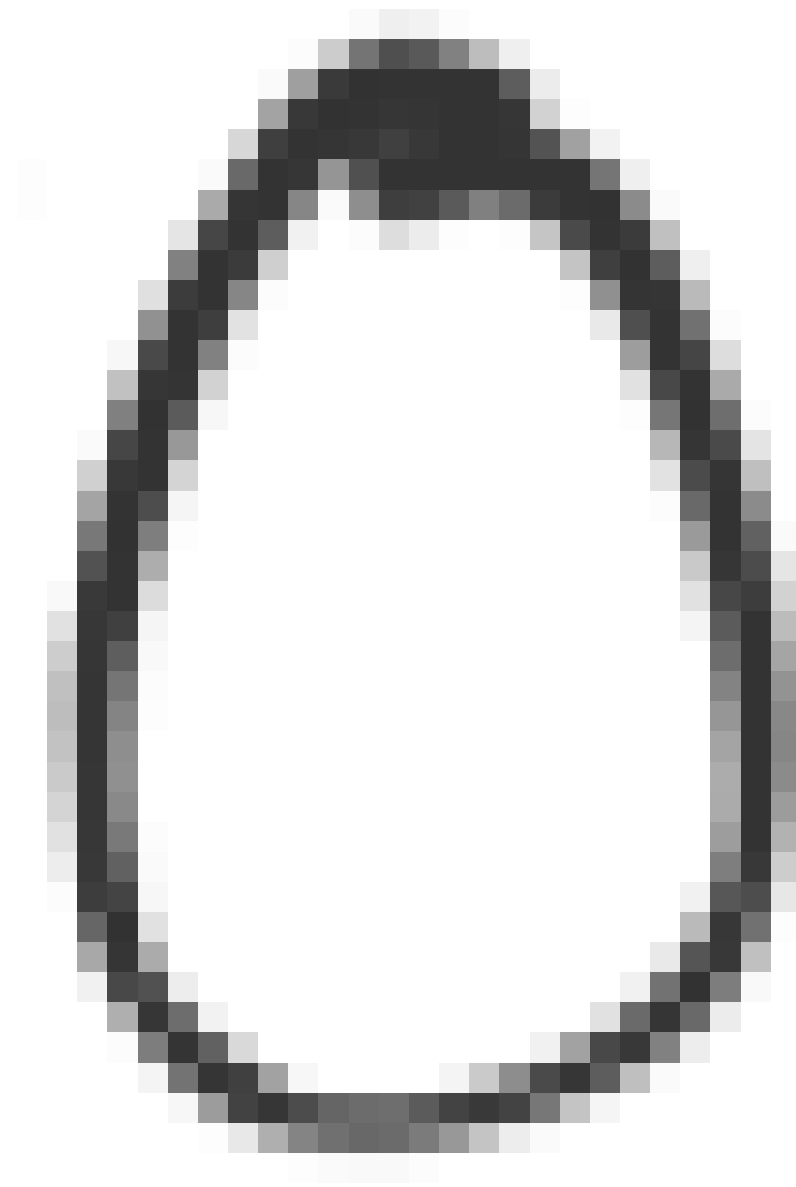
- ◆ **Margin re-scaling loss:** $\psi(t, u) = \max\{0, u - t\}$

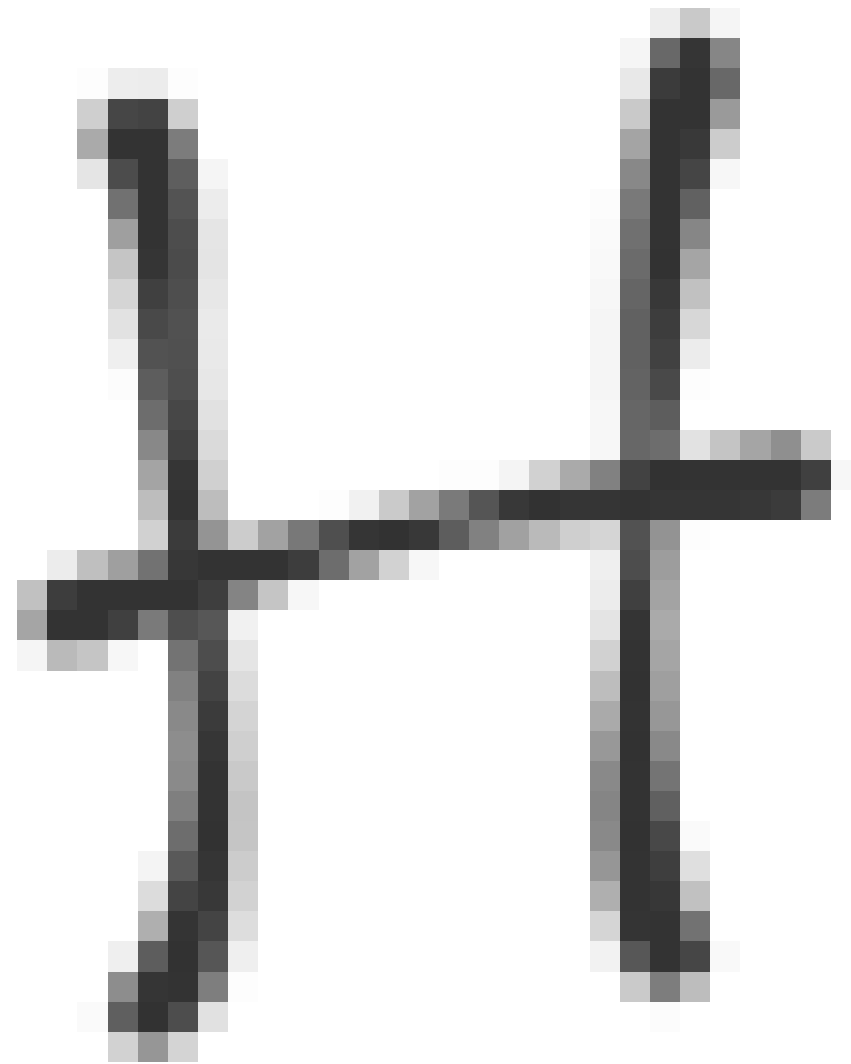
$$\begin{aligned} \ell(h(x), y) &\leq \max_{y' \neq y} \max \left\{ 0, \Delta(y', y) - f(x, y) + f(x, y') \right\} \\ &= \max_{y'} \left(\Delta(y', y) - f(x, y) + f(x, y') \right) \end{aligned}$$

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

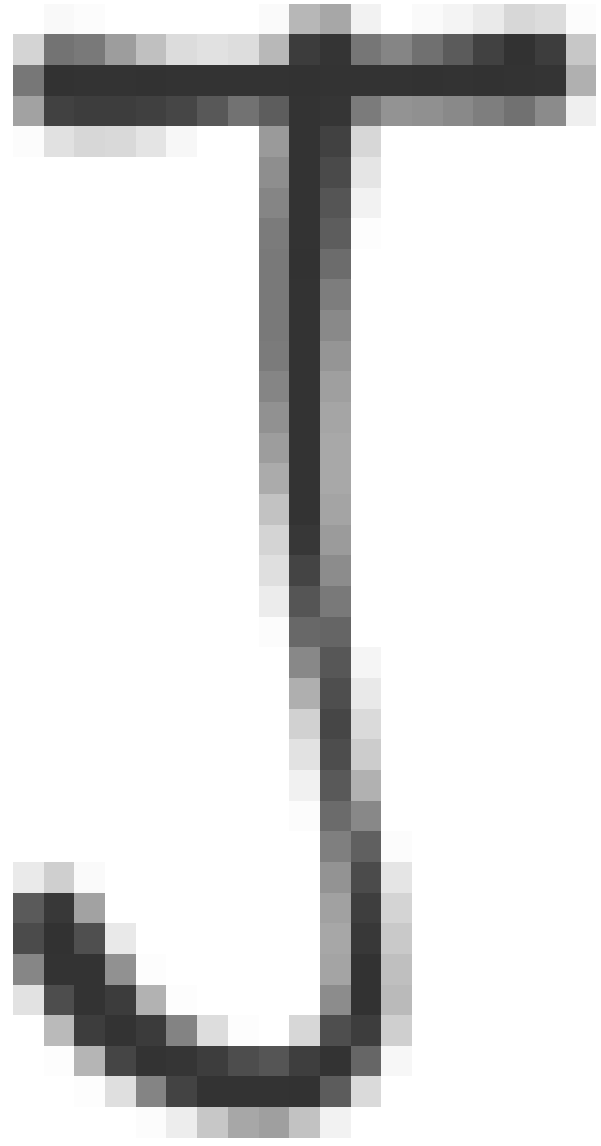
$$\begin{aligned} \ell(h(x), y) &\leq \max_{y' \neq y} \max \left\{ 0, \Delta(y', y) (1 - f(x, y) + f(x, y')) \right\} \\ &= \max_{y'} \Delta(y', y) \left(1 - f(x, y) + f(x, y') \right) \end{aligned}$$

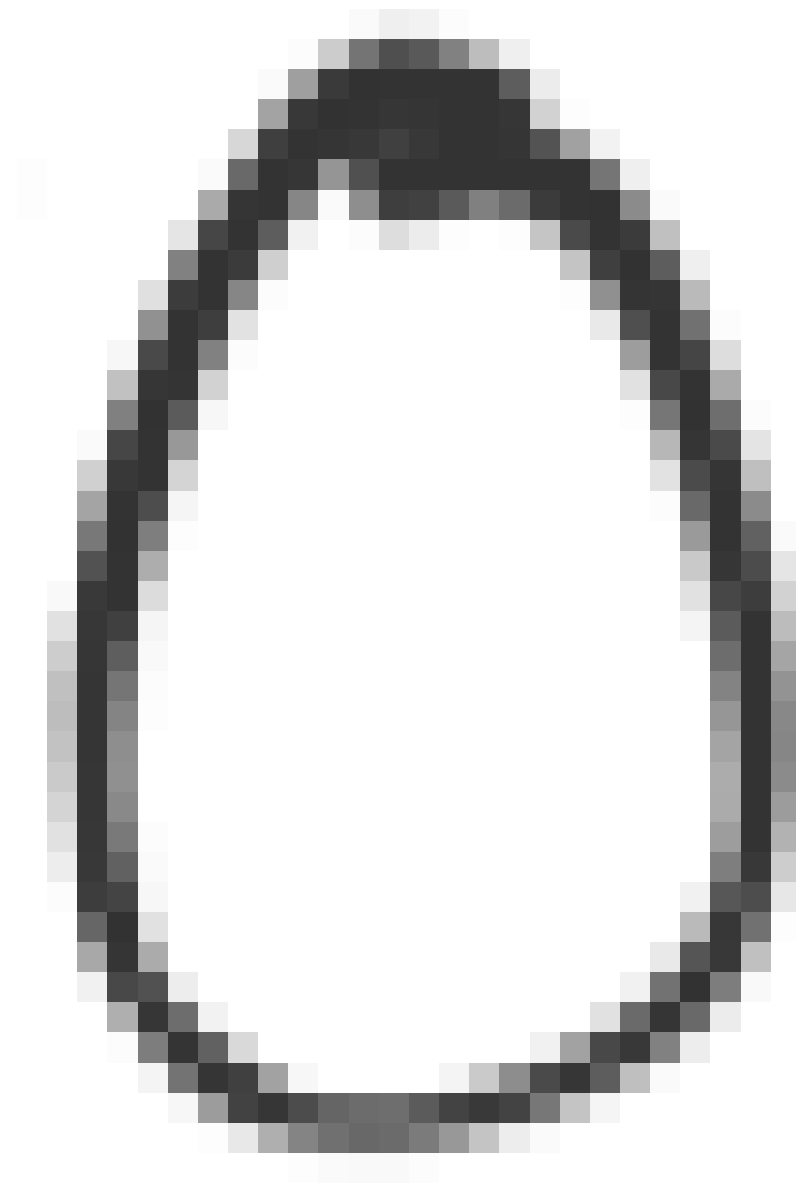


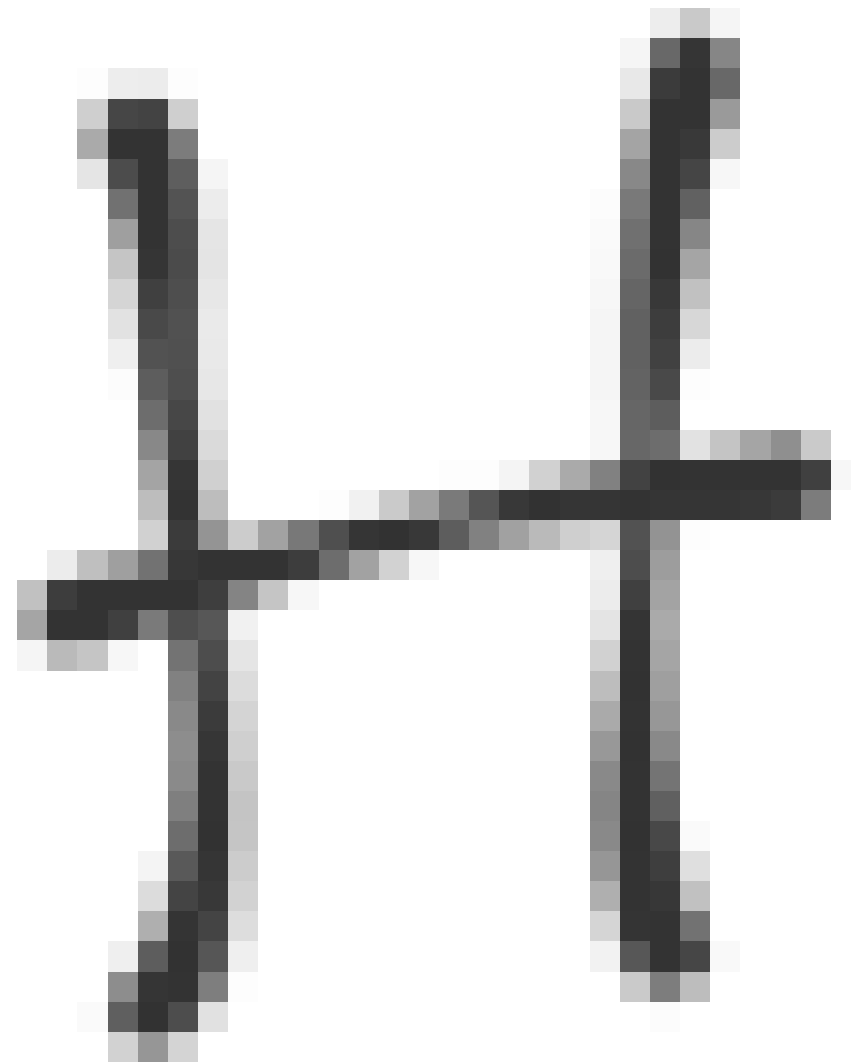




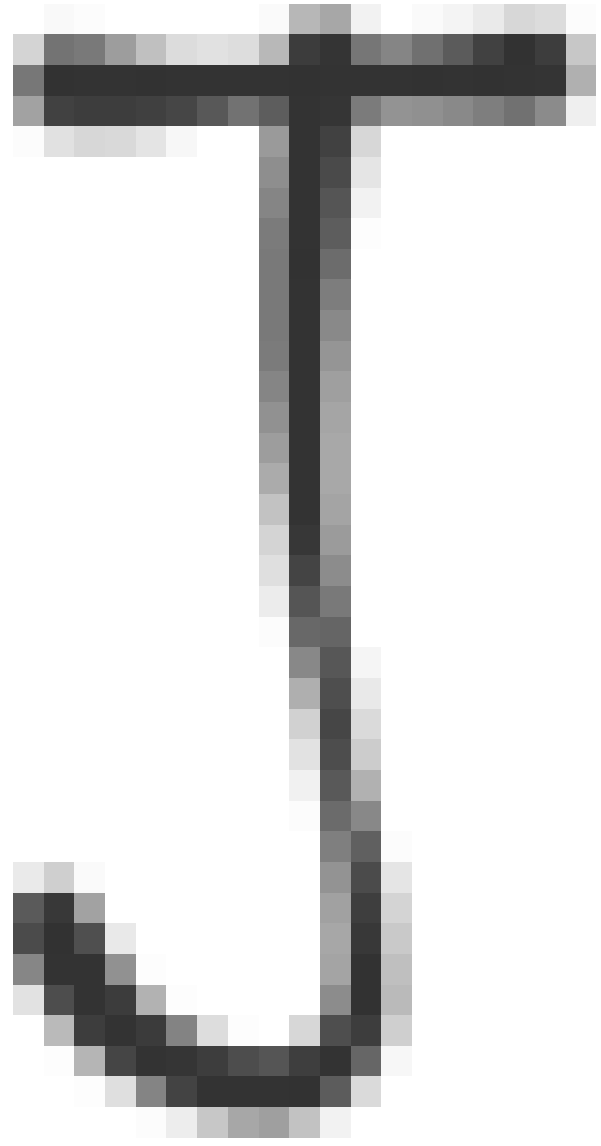
W

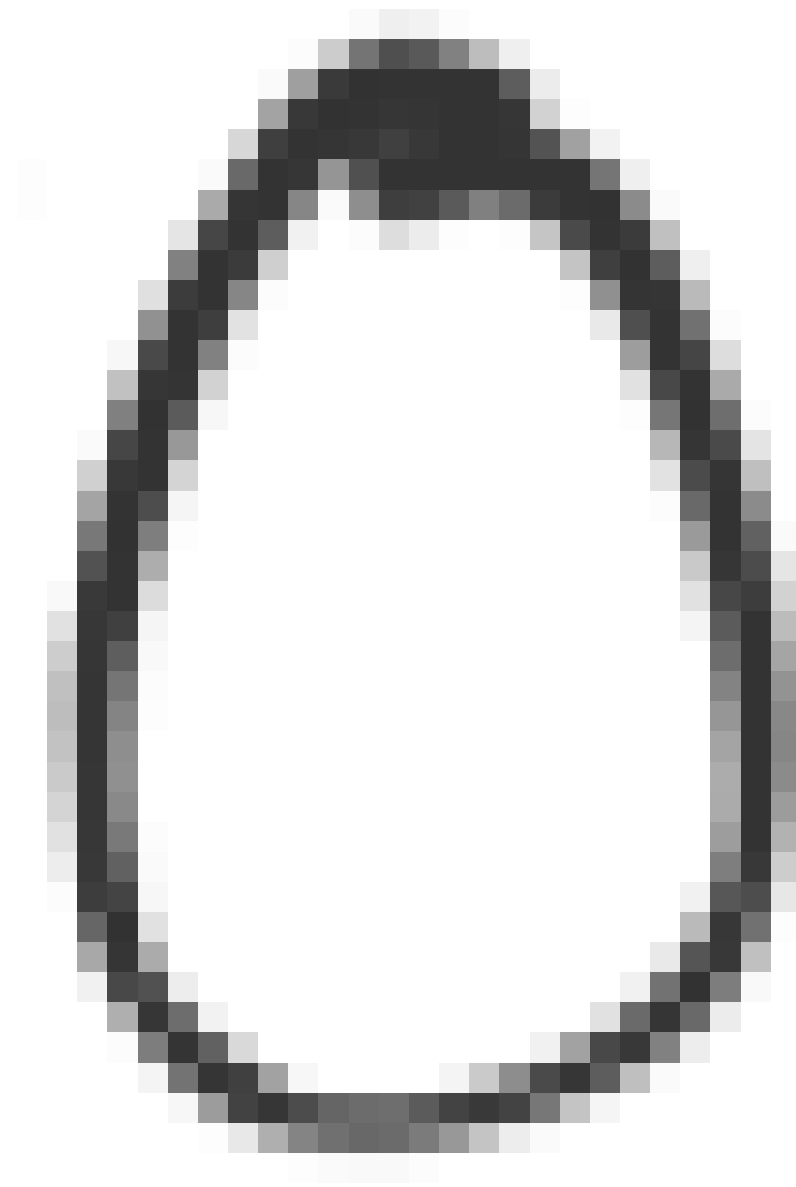


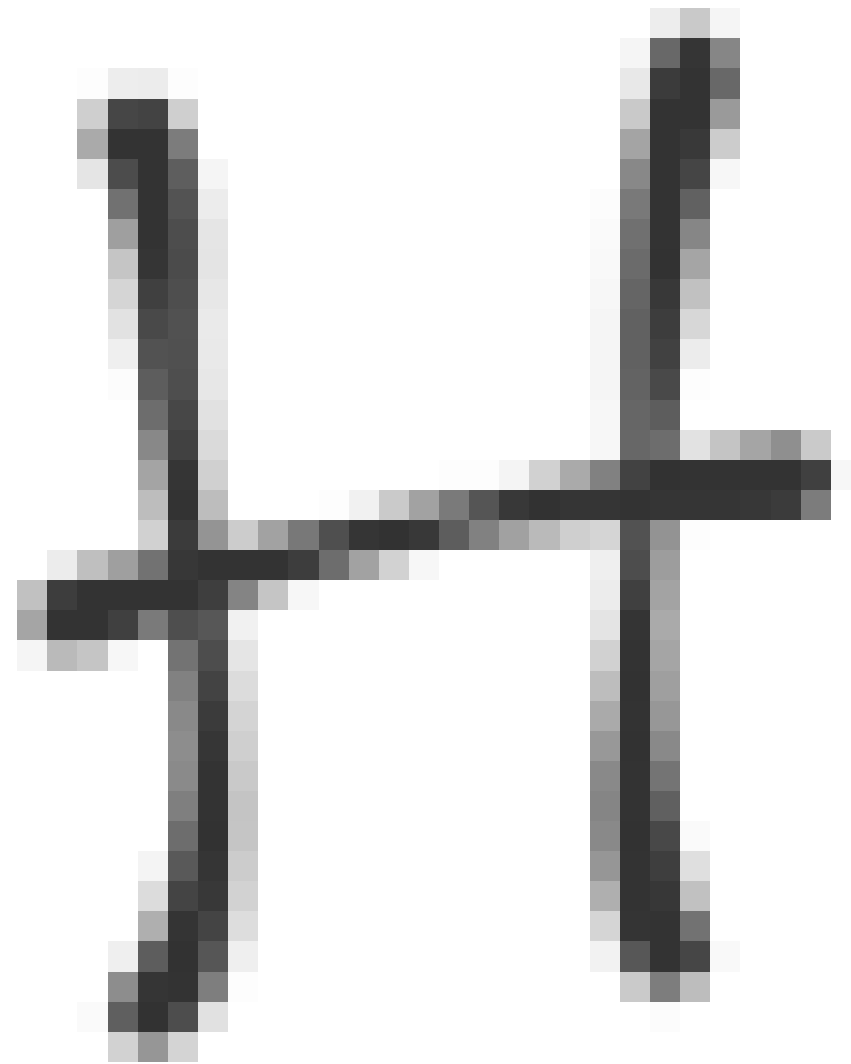




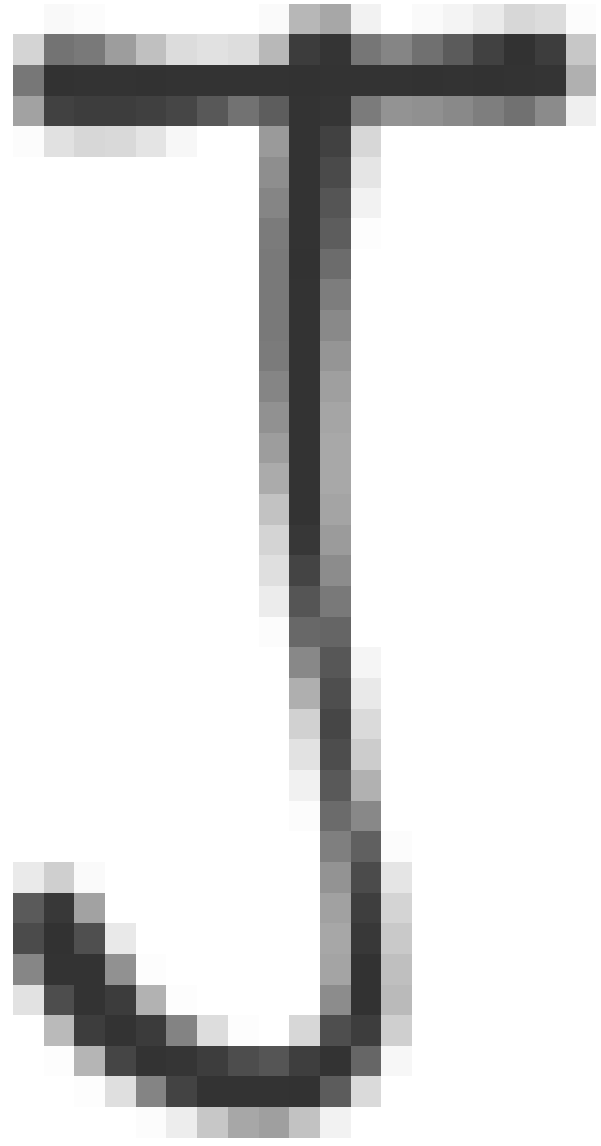
W

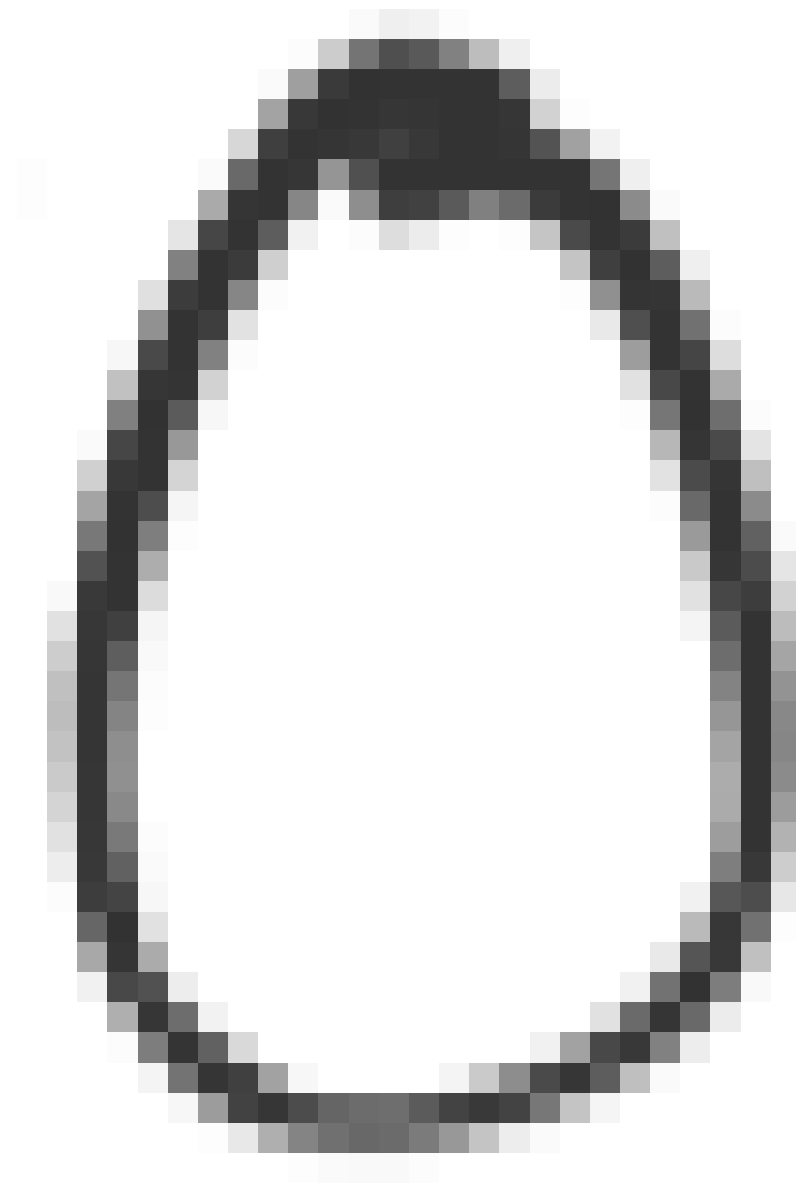


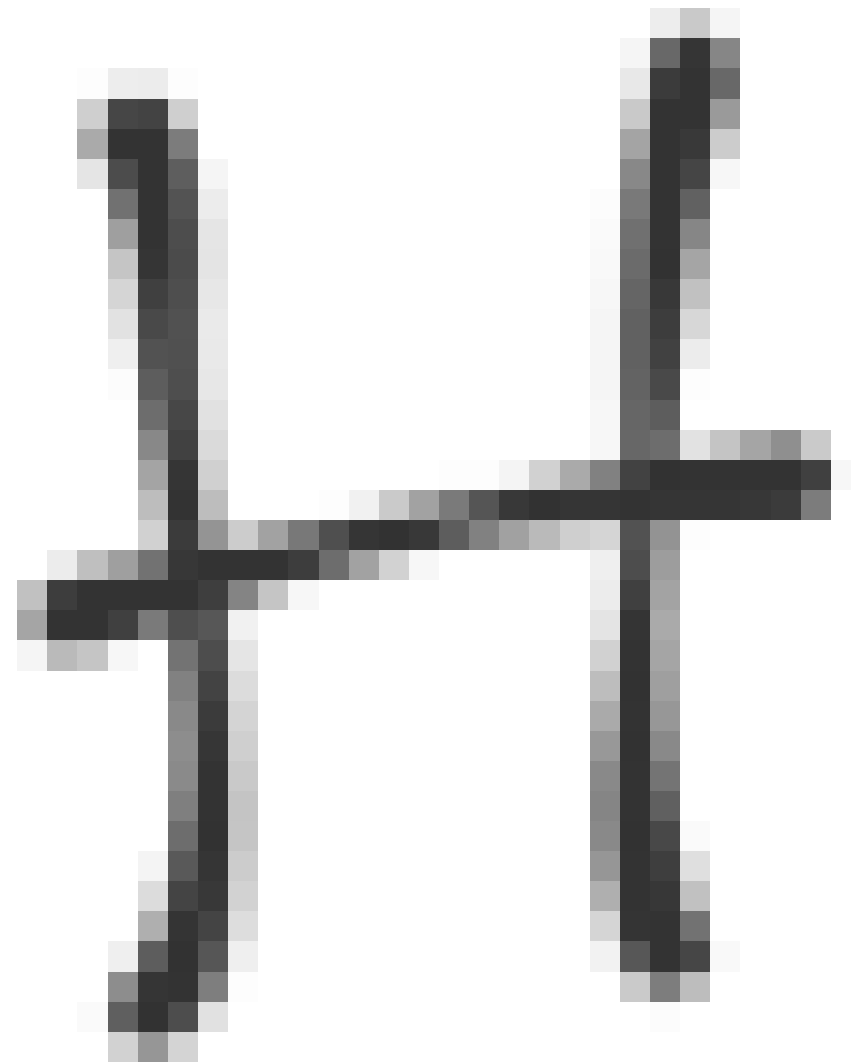




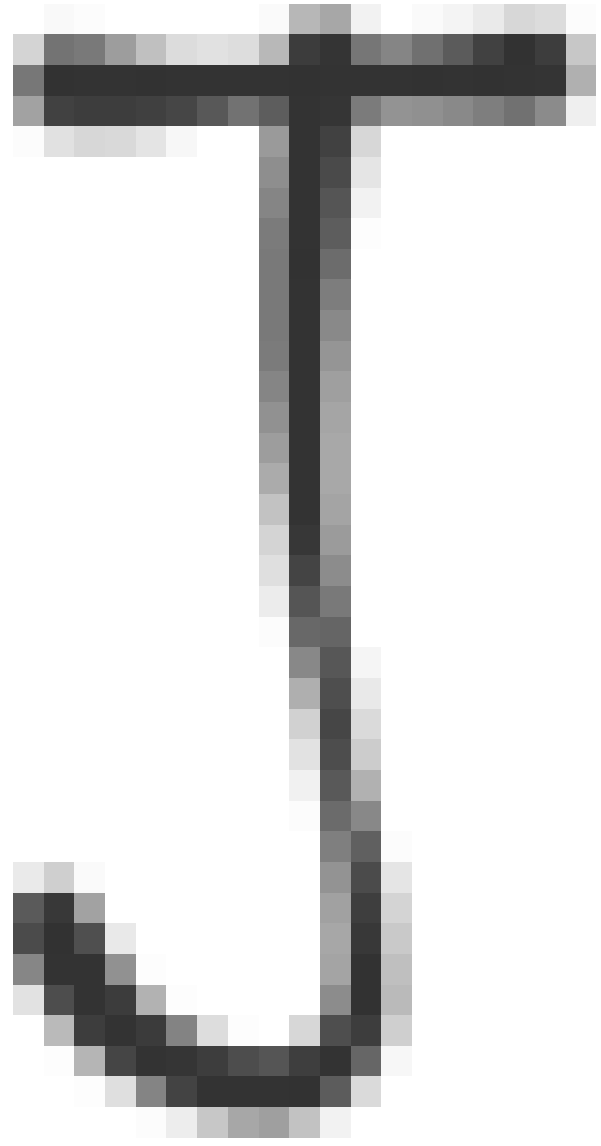
W

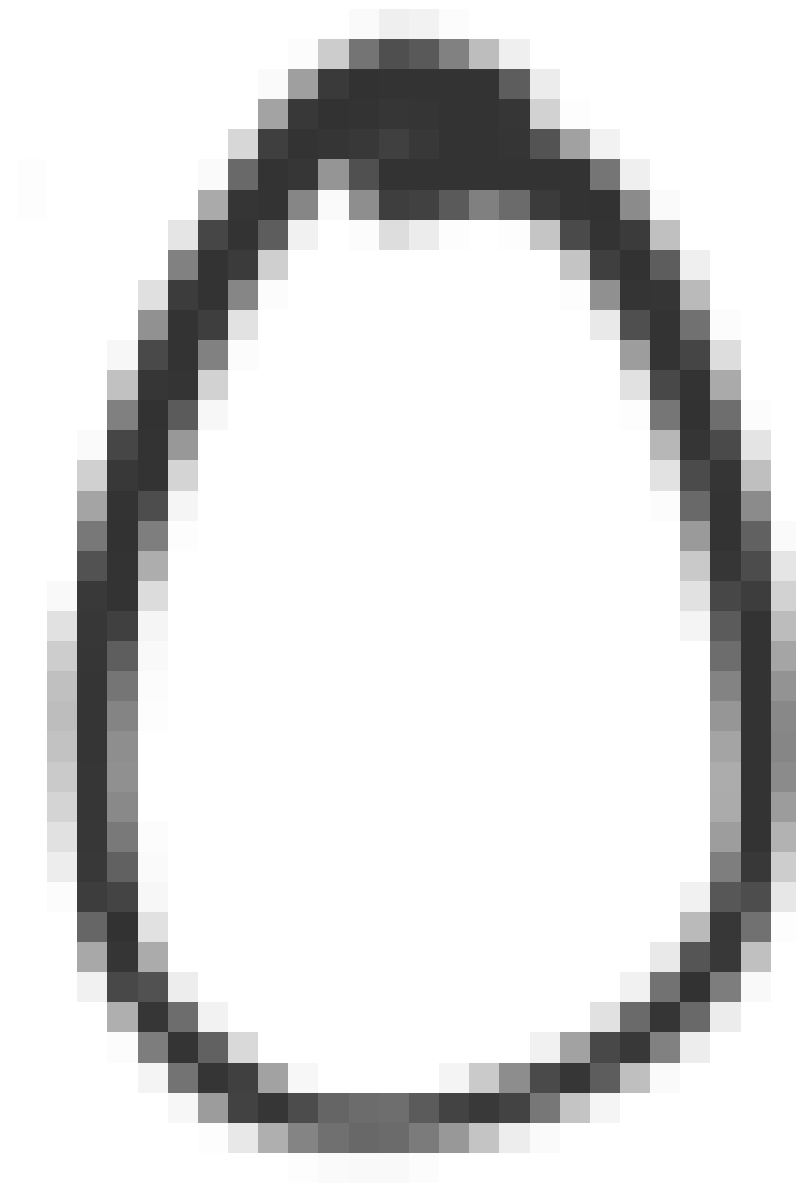


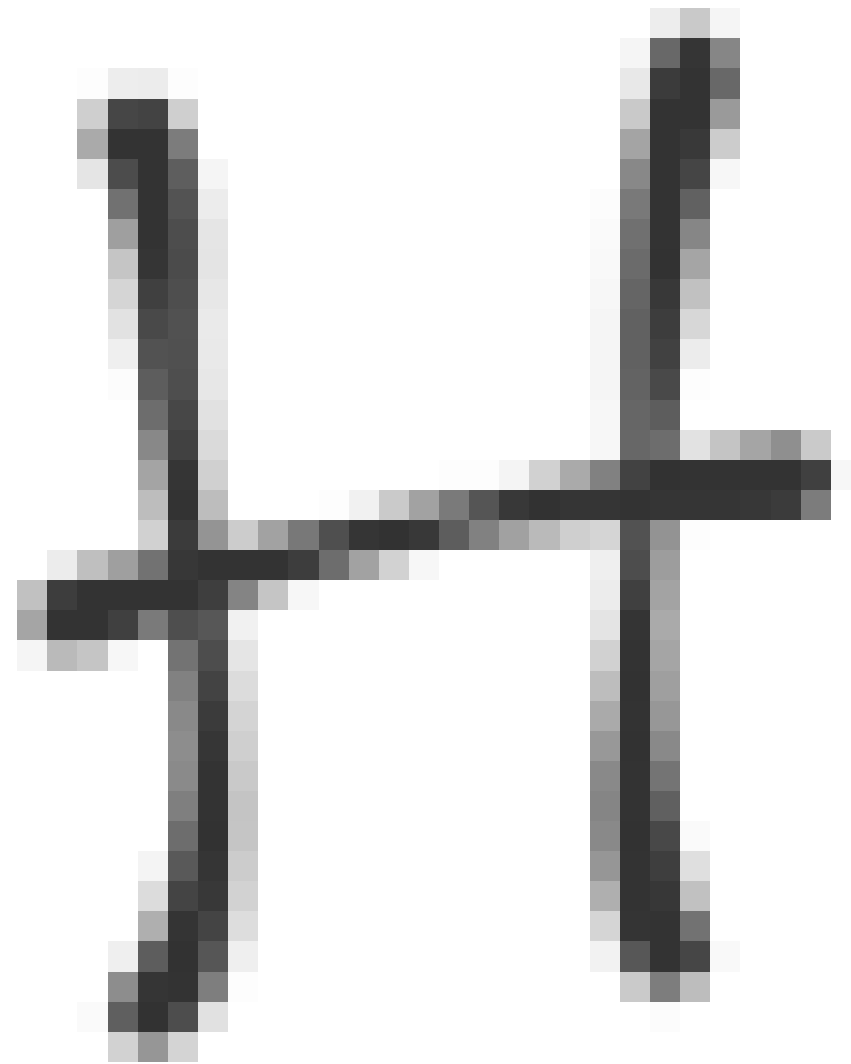




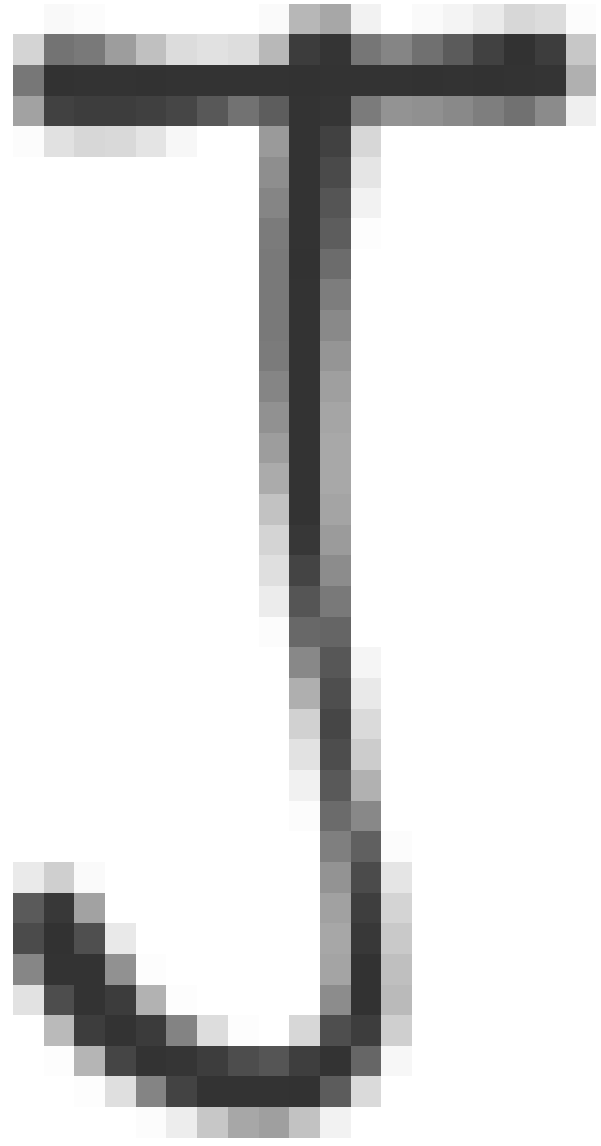
W

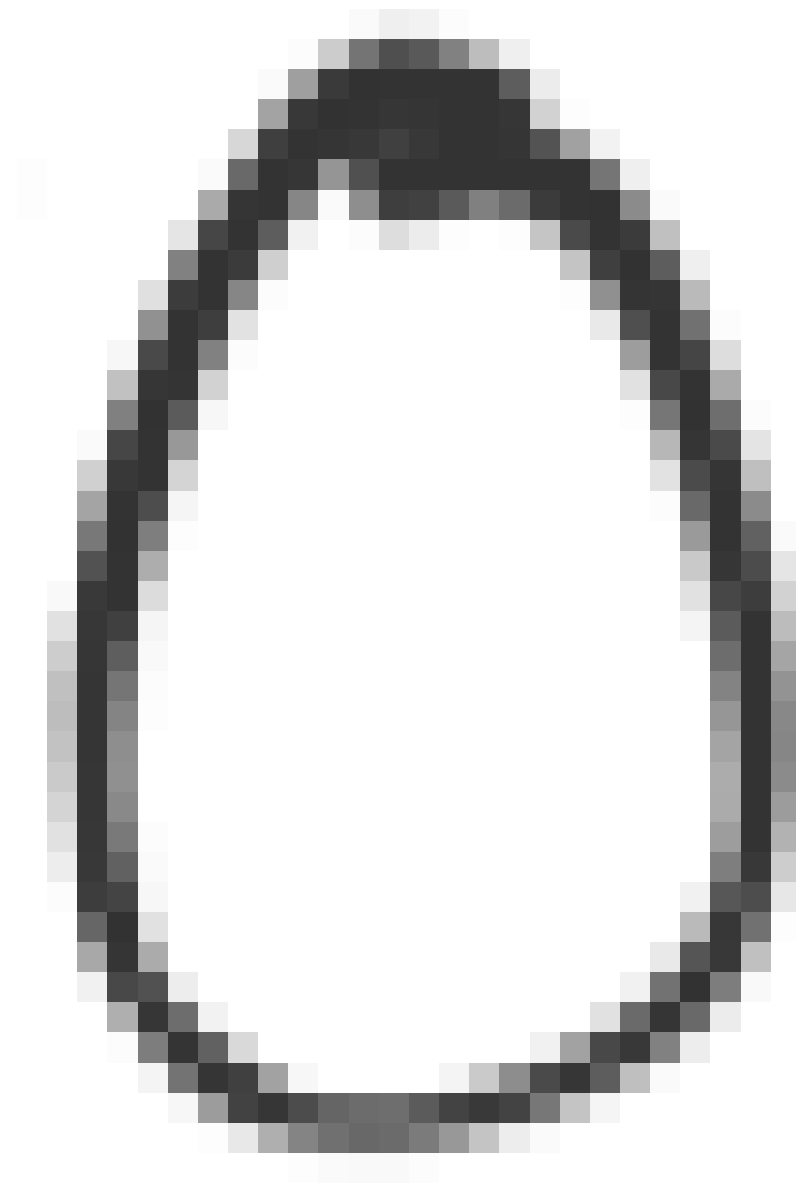


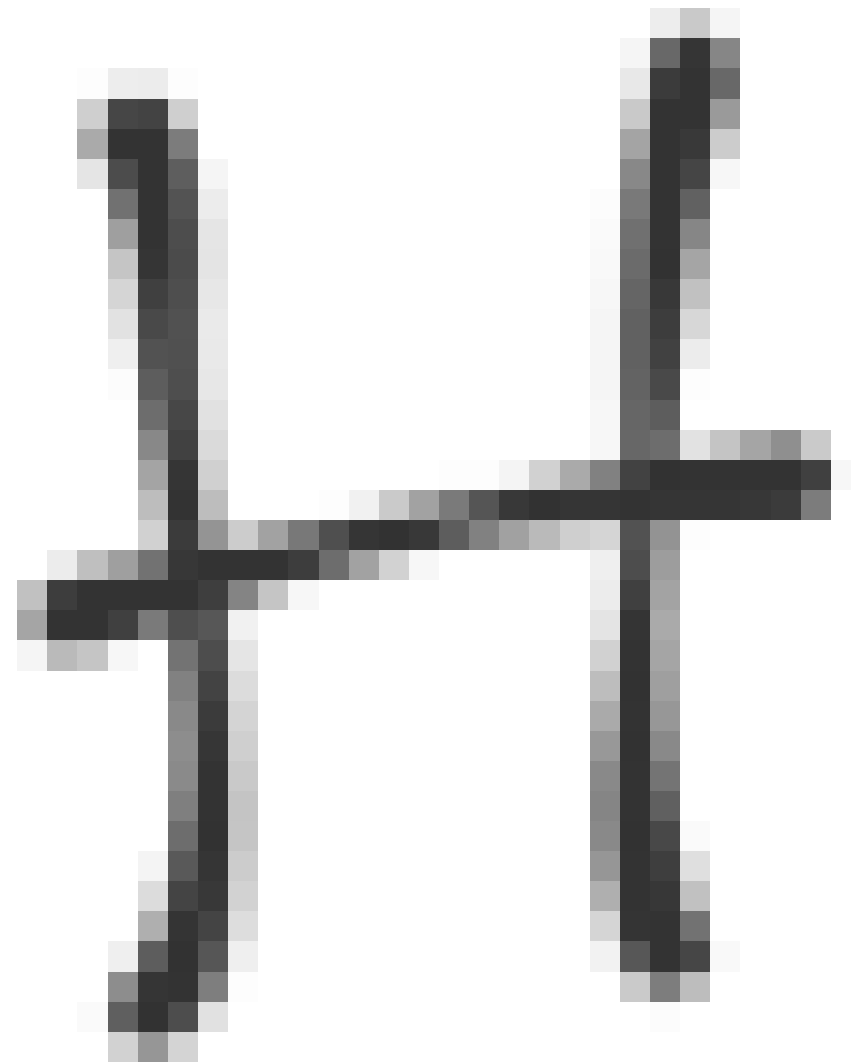




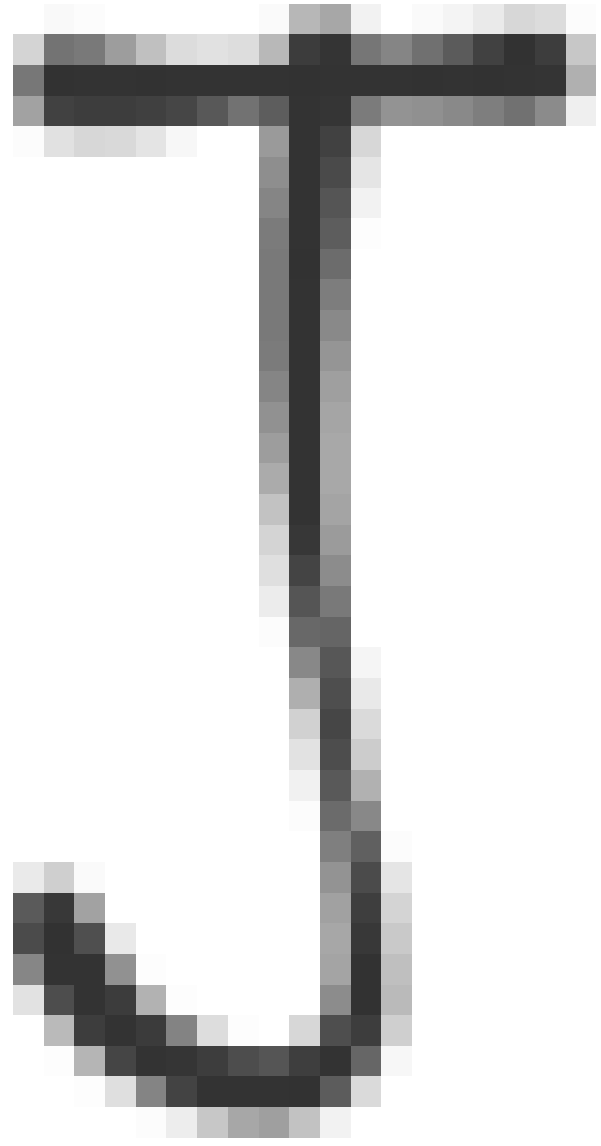
W

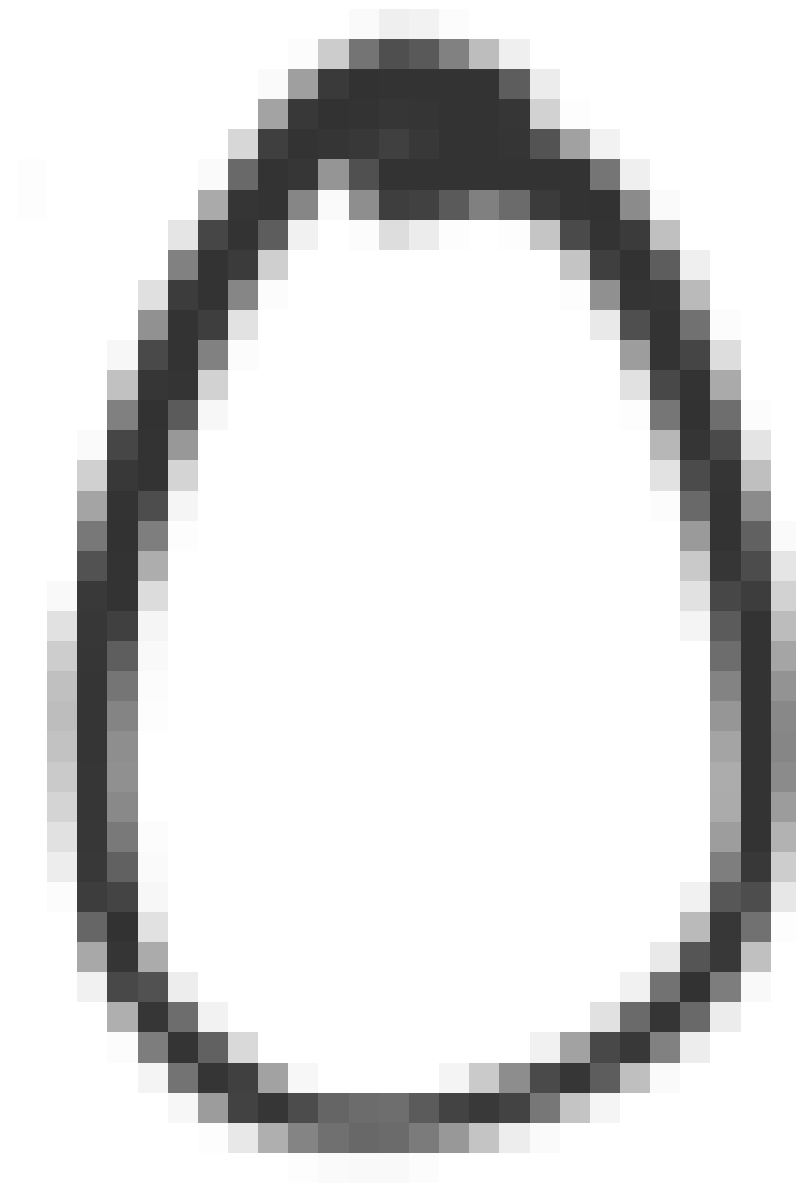


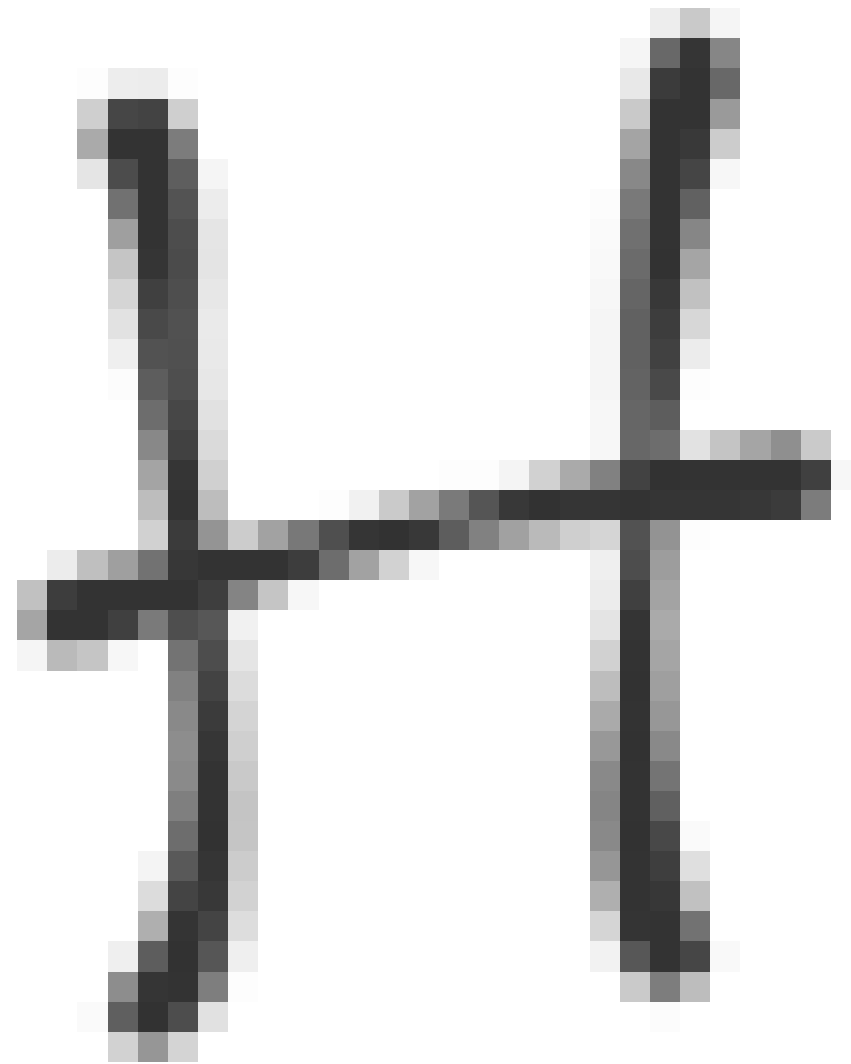




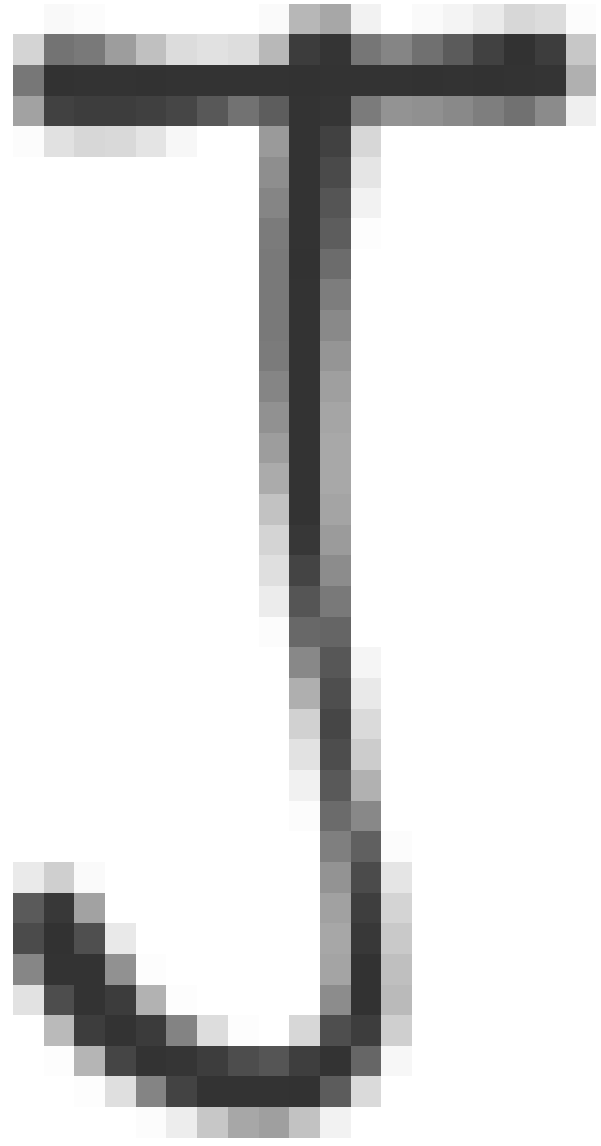
W

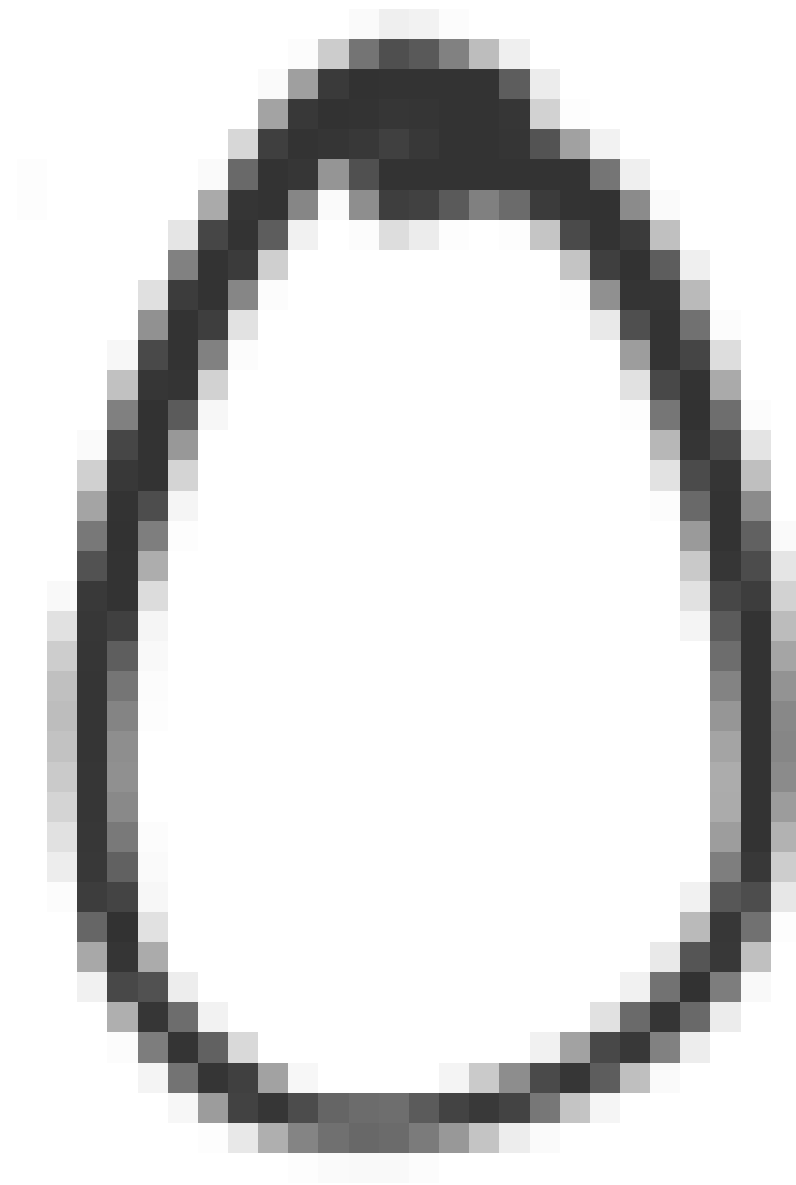


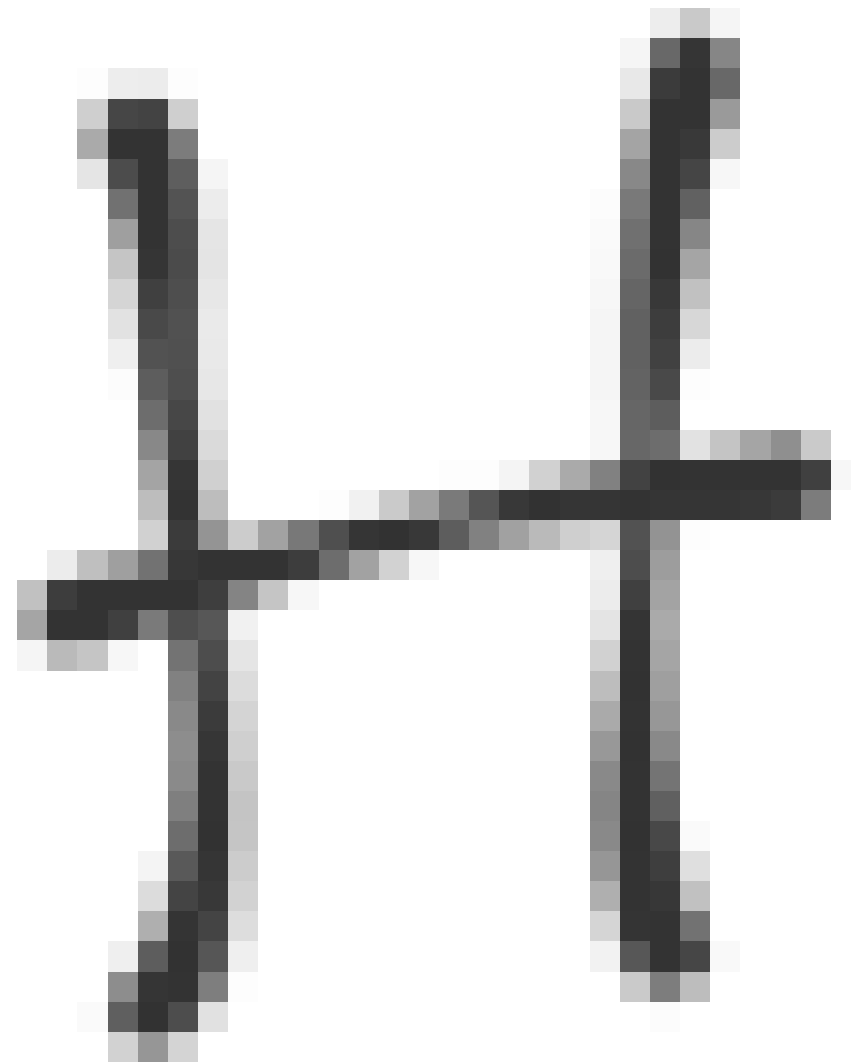




W







W