# Supplementary material for the paper: "What does LIME really see in images?"

June 10, 2021

## Organization of the supplementary material

In this appendix, we present the detailed proof of our main results (Theorem 1 and Proposition 2) and additional qualitative results. We follow the proof scheme of Garreau and von Luxburg [2020]. In a nutshell, when $\lambda = 0$, the main problem

$$\hat{\beta}_n^\lambda \in \mathop{\arg\min}_{\beta \in \mathbb{R}^{d+1}} \left\{ \sum_{i=1}^n \pi_i (y_i - \beta^\top z_i)^2 + \lambda \left\|\beta\right\|^2 \right\} \tag{1}$$

reduces to least squares, with $\hat{\beta}_n$ given in closed-form by

$$\hat{\beta}_n = (Z^\top W Z)^{-1} Z^\top W y\,,$$

with $Z \in \{0,1\}^{n\times d}$ the matrix whose lines are given by the $z_i$s and $W$ the diagonal matrix such that $W_{i,i} = \pi_i$. Setting $\hat{\Sigma}_n := \frac{1}{n} Z^\top W Z$ and $\hat{\Gamma}_n := \frac{1}{n} Z^\top W y$, the study of $\hat{\beta}_n$ can be split in two parts: the examination of $\hat{\Sigma}_n$ (Section 1), and then that of $\hat{\Gamma}_n$ (Section 2). We put everything together in Section 3, proving the concentration of $\hat{\beta}_n$ and providing the expression of $\beta^f$. All technical results are collected in Section 4. Finally, additional qualitative results are presented in Section 5.

## 1 Study of $\hat{\Sigma}_n$

We start by the study of $\hat{\Sigma}_n$, first computing its limit $\Sigma$ when $n \to +\infty$ (Section 1.1). We show that $\Sigma$ is invertible in closed-form in Section 1.2. We then proceed to show that $\hat{\Sigma}_n$ is concentrated around $\Sigma$ in Section 1.3. We conclude this section by obtaining a control on the operator norm of $\Sigma^{-1}$ (Section 1.4), a technical requirement for the proof of the main result.

### 1.1 Computation of $\Sigma$

By definition of $Z$ and $W$, the matrix $\hat{\Sigma}_n$ can be written

$$\hat{\Sigma} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \pi_i & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} & \cdots & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,d} \\ \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} & \cdots & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} z_{i,d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,d} & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} z_{i,d} & \cdots & \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,d} \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}\,.$$

Recall that we defined the random variable $z$ such that $z_i$ is i.i.d. $z$ for any $i$, as well as $\pi$ and $x$ the associated weights and perturbed samples. For any $p \geqslant 0$, we also defined $\alpha_p = \mathbb{E}\left[\pi \prod_{i=1}^p z_i\right]$ (Definition 1). Taking the expectation with respect to $z$ in the previous display, we obtain

$$\Sigma_{j,k} = \begin{cases} \alpha_0 & \text{if } j = k = 0, \\ \alpha_1 & \text{if } j = 0 \text{ and } k > 0 \text{ or } j > 0 \text{ and } k = 0 \text{ or } j = k > 0, \\ \alpha_2 & \text{otherwise.} \end{cases}$$

As promised, it is possible to compute the $\alpha$ coefficients in closed-form. Let us denote by $S$ the number of superpixel deletions. Since the coordinates of $z$ are i.i.d. Bernoulli with parameter $1/2$, we deduce that $S$ is a *binomial* random variable of parameters $d$ and $1/2$. Note that, conditionally to $S = s$, $\sum_j z_j = d - s$ and therefore $\pi = \psi(s/d)$ with

$$\forall t \in [0,1], \quad \psi(t) := \exp\left(\frac{-(1-\sqrt{1-t})^2}{2\nu^2}\right) \tag{2}$$

as in the paper. As a consequence of these observations, we have:

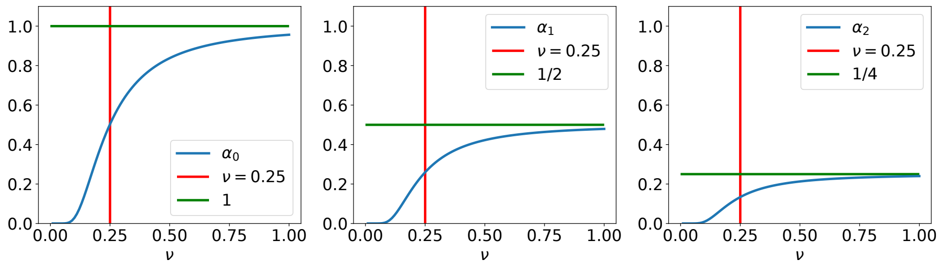


Figure 1: The first three $\alpha$ coefficients as a function of the bandwidth $\nu$ for $d = 50$. In green the limit value given by Lemma 1.

**Proposition 1 (Computation of the $\alpha$ coefficients).** *Let $p \geqslant 0$ be an integer. Then*

$$\alpha_p = \frac{1}{2^d}\sum_{s=0}^{d}\binom{d-p}{s}\psi(s/d)\,.$$

*Proof.* We write

$$\begin{aligned}
\alpha_p &= \mathbb{E}\left[\pi z_1 \cdots z_p\right] \\
&= \sum_{s=0}^{d} \mathbb{E}_s\left[\pi z_1 \cdots z_p\right]\mathbb{P}\left(S = s\right) && \text{(law of total expectation)} \\
&= \frac{1}{2^d}\sum_{s=0}^{d}\binom{d}{s}\mathbb{E}_s\left[\pi z_1 \cdots z_p | z_1 = 1, \ldots, z_p = 1\right]\mathbb{P}_s\left(z_1 = 1, \ldots, z_p = 1\right) && (S \sim \mathcal{B}(n, 1/2)) \\
&= \frac{1}{2^d}\sum_{s=0}^{d}\binom{d}{s}\psi(s/d)\mathbb{P}_s\left(z_1 = 1, \ldots, z_p = 1\right) && \text{(definition of } \psi) \\
\alpha_p &= \frac{1}{2^d}\sum_{s=0}^{d}\binom{d}{s}\frac{(d-p)!}{d!}\cdot\frac{(d-s)!}{(d-s-p)!}\psi(s/d) && \text{(Lemma 3)}
\end{aligned}$$

We conclude by some algebra. □

It is quite straightforward to compute the limits of the $\alpha$ coefficients when $\nu \to +\infty$. In fact, since $\mathrm{e}^{-1/(2\nu^2)} \leqslant \psi(t) \leqslant 1$ for any $\nu > 0$, we have the following bounds on $\alpha_p$:

**Lemma 1 (Bounding the $\alpha$ coefficients).** *For any $p \geqslant 0$, we have*

$$\frac{\mathrm{e}^{\frac{-1}{2\nu^2}}}{2^p} \leqslant \alpha_p \leqslant \frac{1}{2^p}\,.$$

*In particular, when $\nu \to +\infty$, we have $\alpha_p \to \frac{1}{2^p}$ for any $p \geqslant 0$.*

We demonstrate these approximations in Figure 1.

## 1.2 $\sigma$ coefficients

Since the structure of $\Sigma$ is the same as in the text case [Mardaoui and Garreau, 2021], we can invert it similarly.

**Proposition 2 (Inverse of $\Sigma$).** *For any $d \geqslant 1$, recall that we defined*

$$\begin{cases}
\sigma_1 &= -\alpha_1\,, \\
\sigma_2 &= \frac{(d-2)\alpha_0\alpha_2 - (d-1)\alpha_1^2 + \alpha_0\alpha_1}{\alpha_1 - \alpha_2}\,, \\
\sigma_3 &= \frac{\alpha_1^2 - \alpha_0\alpha_2}{\alpha_1 - \alpha_2}\,,
\end{cases}$$

*and $c_d = (d-1)\alpha_0\alpha_2 - d\alpha_1^2 + \alpha_0\alpha_1$. Let us further define $\sigma_0 := (d-1)\alpha_2 + \alpha_1$. Assume that $c_d \neq 0$ and $\alpha_1 \neq \alpha_2$. Then $\Sigma$ is invertible, and it holds that*

$$\Sigma^{-1} = \frac{1}{c_d}\begin{pmatrix}
\sigma_0 & \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\
\sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_3 \\
\sigma_1 & \sigma_3 & \sigma_2 & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \sigma_3 \\
\sigma_1 & \sigma_3 & \cdots & \sigma_3 & \sigma_2
\end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}\,. \tag{3}$$

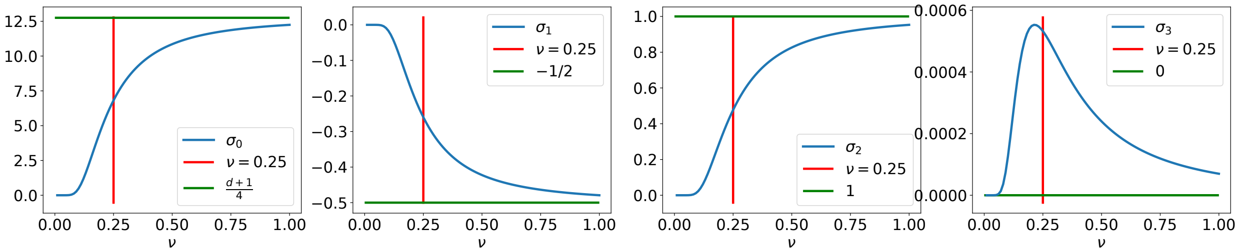


Figure 2: The first four $\sigma$ coefficients as a function of the bandwidth $\nu$ for $d = 50$. In green, the limit values given by Eq. (4).

From Lemma 1, we deduce

$$\sigma_0 \to \frac{d+1}{4}\,, \quad \sigma_1 \to \frac{-1}{2}\,, \quad \sigma_2 \to 1\,, \quad \sigma_3 \to 0\,, \quad \text{and} \quad c_d \to 1/4\,. \tag{4}$$

when $\nu \to +\infty$. We illustrate this in Figure 2. Now, Proposition 2 requires $\alpha_1 \neq \alpha_2$ and $c_d \neq 0$ in order for $\Sigma$ to be invertible. One of the consequences of the following result is that these conditions are always satisfied.

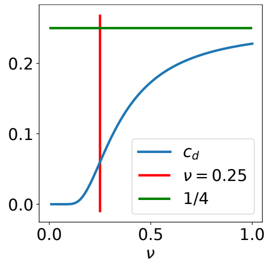


Figure 3: Evolution of $c_d$ with respect to $\nu$ when $d = 50$.

**Proposition 3 ($\Sigma$ is invertible).** *Let $d \geqslant 1$ and $\nu > 0$. Then $\alpha_1 - \alpha_2 \geqslant \mathrm{e}^{\frac{-1}{2\nu^2}}/4$ and $c_d \geqslant \mathrm{e}^{\frac{-1}{\nu^2}}/4$.*

Note that in this case the lower bound obtained on $c_d$ is tight. We show the evolution of $c_d$ with respect to the bandwidth in Figure 3.

*Proof.* By definition of the $\alpha$ coefficients and Pascal identity, it holds that

$$\alpha_p - \alpha_{p+1} = \frac{1}{2^d} \sum_{s=0}^{d} \binom{d-p-1}{s-1} \psi\left(\frac{s}{d}\right)\,, \tag{5}$$

for any $p \geqslant 0$. Since $\mathrm{e}^{-1/(2\nu^2)} \leqslant \psi(t) \leqslant 1$ for any $1 \leqslant t \leqslant 1$, we deduce from Eq. (5) that, for any $p \geqslant 0$,

$$\frac{\mathrm{e}^{\frac{-1}{2\nu^2}}}{2^{p+1}} \leqslant \alpha_p - \alpha_{p+1} \leqslant \frac{1}{2^{p+1}}\,. \tag{6}$$

We deduce the lower bound on $\alpha_1 - \alpha_2$ by setting $p = 1$ in the previous display.

Let us turn to $c_d$. We write

$$\begin{aligned} c_d &= d\alpha_1(\alpha_0 - \alpha_1) - (d-1)\alpha_0(\alpha_1 - \alpha_2) \\ &= \frac{1}{4^d}\left[ d \cdot \sum_{s=0}^{d} \binom{d-1}{s} \psi\left(\frac{s}{d}\right) \cdot \sum_{s=0}^{d} \binom{d-1}{s-1} \psi\left(\frac{s}{d}\right) - (d-1) \cdot \sum_{s=0}^{d} \binom{d}{s} \psi\left(\frac{s}{d}\right) \cdot \sum_{s=0}^{d} \binom{d-2}{s-1} \psi\left(\frac{s}{d}\right) \right] \\ & \qquad\qquad \text{(using Eq. (5))} \\ c_d &= \frac{1}{4^d}\left[ \sum_{s=0}^{d} \binom{d-1}{s} \psi\left(\frac{s}{d}\right) \cdot \sum_{s=0}^{d} s\binom{d}{s} \psi\left(\frac{s}{d}\right) - \sum_{s=0}^{d} \binom{d}{s} \psi\left(\frac{s}{d}\right) \cdot \sum_{s=0}^{d} s\binom{d-1}{s} \psi\left(\frac{s}{d}\right) \right]\,, \end{aligned}$$

where we used elementary properties of the binomial coefficients in the last display. For any $0 \leqslant s \leqslant d$, let us set

$$A_s := \binom{d-1}{s}\sqrt{\psi\left(\frac{s}{d}\right)}\,, B_s := s\sqrt{\psi\left(\frac{s}{d}\right)}\,, C_s := \sqrt{\psi\left(\frac{s}{d}\right)}\,, \text{ and } D_s := \binom{d}{s}\sqrt{\psi\left(\frac{s}{d}\right)}\,.$$

With these notation,

$$c_d = \frac{1}{4^d}\left[ \sum_s A_s C_s \cdot \sum_s B_s D_s - \sum_s A_s B_s \cdot \sum C_s D_s \right]\,.$$

According to the four-letter identity (Proposition 13), we can rewrite $c_d$ as

$$\begin{aligned} c_d &= \frac{1}{4^d} \sum_{s<t} (A_s D_t - A_t D_s)(C_s B_t - C_t B_s) \\ &= \frac{1}{4^d} \sum_{s<t} (t-s) \left( \binom{d-1}{s}\binom{d}{t} - \binom{d-1}{t}\binom{d}{s} \right) \psi\left(\frac{s}{d}\right) \psi\left(\frac{t}{d}\right) \\ c_d &= \frac{1}{d \cdot 4^d} \sum_{s<t} \binom{d}{s}\binom{d}{t} (s-t)^2 \psi\left(\frac{s}{d}\right) \psi\left(\frac{t}{d}\right)\,. \end{aligned}$$

Since $\mathrm{e}^{-1/(2\nu^2)} \leqslant \psi(t) \leqslant 1$ for any $1 \leqslant t \leqslant 1$, all that is left to do is to control the double sum. According to Proposition 14, we have

$$\sum_{s<t} \binom{d}{s}\binom{d}{t}(s-t)^2 = d \cdot 4^{d-1} \, .$$

We deduce that

$$\frac{\mathrm{e}^{\frac{-1}{2\nu^2}}}{4} \leqslant c_d \leqslant \frac{1}{4} \, . \tag{7}$$

□

We conclude this section with useful relationships between $\alpha$ and $\sigma$ coefficients.

**Proposition 4 (Useful equalities).** *Let $\alpha_p$, $\sigma_p$, and $c_d$ be defined as above. Then it holds that*

$$\sigma_0\alpha_1 + \sigma_1\alpha_1 + (d-1)\sigma_1\alpha_2 = 0 \, , \tag{8}$$

$$\sigma_1\alpha_1 + \sigma_2\alpha_1 + (d-1)\sigma_3\alpha_2 = c_d \, , \tag{9}$$

$$\sigma_1\alpha_1 + \sigma_2\alpha_2 + \sigma_3\alpha_1 + (d-2)\sigma_3\alpha_2 = 0 \, , \tag{10}$$

$$\sigma_1\alpha_0 + \sigma_2\alpha_1 + (d-1)\sigma_3\alpha_1 = 0 \, , \tag{11}$$

$$\sigma_0\alpha_0 + d\sigma_1\alpha_1 = c_d \, . \tag{12}$$

*Proof.* Straightforward from the definitions. □

## 1.3 Concentration of $\hat{\Sigma}_n$

We now turn to the concentration of $\hat{\Sigma}_n$ around $\Sigma$. More precisely, we show that $\hat{\Sigma}_n$ is close to $\Sigma$ in operator norm, with high probability. Since the definition of $\hat{\Sigma}_n$ is identical to the one in the Tabular LIME case, we can use the proof machinery of Garreau and von Luxburg [2020].

**Proposition 5 (Concentration of $\hat{\Sigma}_n$).** *For any $t \geqslant 0$,*

$$\mathbb{P}\left( \left\| \hat{\Sigma}_n - \Sigma \right\|_{\mathrm{op}} \geqslant t \right) \leqslant 4d \cdot \exp\left( \frac{-nt^2}{32d^2} \right) \, .$$

*Proof.* We can write $\hat{\Sigma} = \frac{1}{n}\sum_i \pi_i Z_i Z_i^\top$. The summands are bounded i.i.d. random variables, thus we can apply the matrix version of Hoeffding inequality. More precisely, the entries of $\hat{\Sigma}_n$ belong to $[0,1]$ by construction, and Lemma 1 guarantees that the entries of $\Sigma$ also belong to $[0,1]$. Therefore, if we set $M_i := \frac{1}{n}\pi_i Z_i Z_i^\top - \Sigma$, then the $M_i$ satisfy the assumptions of Theorem 21 in Garreau and von Luxburg [2020] and we can conclude since $\frac{1}{n}\sum_i M_i = \hat{\Sigma}_n - \Sigma$. □

## 1.4 Control of $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}$

In this section, we obtain a control on the operator norm of the inverse covariance matrix. Our strategy is to bound the norm of the $\sigma$ coefficients. We start with the control of $\alpha_1^2 - \alpha_0\alpha_2$, a quantity appearing in $\sigma_2$ and $\sigma_3$.

**Lemma 2 (Control of $\alpha_1^2 - \alpha_0\alpha_2$).** *For any $d \geqslant 2$, we have*

$$\left|\alpha_1^2 - \alpha_0\alpha_2\right| \leqslant \frac{1}{2d} \, .$$

*Proof.* By definition of the $\alpha$ coefficients, we know that

$$\alpha_1^2 - \alpha_0\alpha_2 = \frac{1}{4^d}\left[ \left( \sum_{s=0}^{d} \binom{d-1}{s} \psi\left(\frac{s}{d}\right) \right)^2 - \left( \sum_{s=0}^{d} \binom{d}{s} \psi\left(\frac{s}{d}\right) \right) \cdot \left( \sum_{s=0}^{d} \binom{d-2}{s} \psi\left(\frac{s}{d}\right) \right) \right] \, .$$

Let us ignore the $1/4^d$ normalization for now, and set $w_s := \binom{d}{s}\psi\left(\frac{s}{d}\right)$. Elementary manipulations of the binomial coefficients allow us to rewrite the previous display as

$$\left( \sum_{s=0}^{d} \frac{d-s}{d} w_s \right)^2 - \left( \sum_{s=0}^{d} w_s \right) \cdot \left( \sum_{s=0}^{d} \frac{d-s}{d} \cdot \frac{d-s-1}{d-1} w_s \right) \, . \tag{13}$$

Let us notice that

$$\frac{d-s}{d} - \frac{d-s-1}{d-1} = \frac{s}{d(d-1)} \, .$$

Thus we can split Eq. (13) in two parts.

The first part is reminiscent of the Cauchy-Schwarz-like expression that appears in the proof of Proposition 3:

$$\left(\sum_{s=0}^{d} \frac{d-s}{d} w_s\right)^2 - \left(\sum_{s=0}^{d} w_s\right) \cdot \left(\sum_{s=0}^{d} \frac{(d-s)^2}{d^2} w_s\right) . \tag{14}$$

We use, again, the four letter identity (Proposition 13) to bound this term. Namely, for any $0 \leqslant s \leqslant d$, let us set

$$A_s = B_s := \frac{d-s}{d}\sqrt{w_s}\,, \qquad \text{and} \qquad C_s = D_s := \sqrt{w_s}\,.$$

Then we can rewrite Eq. (14) as

$$\sum_{s<t} (A_s D_t - A_t D_s)(C_s B_t - C_t B_s) = \frac{-1}{d^2} \sum_{s<t} (t-s)^2 \binom{d}{s}\binom{d}{t} \psi\left(\frac{s}{d}\right) \psi\left(\frac{t}{d}\right) . \tag{15}$$

According to Proposition 14, Eq. (15) is bounded by $d \cdot 4^{d-1}/d^2 = 4^{d-1}/d$.

The second part of Eq. (13) reads

$$\left(\sum_{s=0}^{d} w_s\right) \cdot \left(\sum_{s=0}^{d} \frac{d-s}{d} \cdot \frac{s}{d(d-1)} w_s\right) .$$

Since $\psi$ is bounded by 1, coming back to the definition of the $w_s$, it is straightforward to show that $\left|\sum_s w_s\right| \leqslant 2^d$ and that $\left|\sum_s s(d-s) w_s\right| \leqslant d(d-1)2^{d-2}$. We deduce that (the absolute value of) this second term is upper bounded by $4^{d-1}/d$.

Putting together the bounds obtained on both terms and renormalizing by $4^d$, we obtain that

$$\left|\alpha_1^2 - \alpha_0\alpha_2\right| \leqslant \frac{1}{4^d}\left[\frac{4^{d-1}}{d} + \frac{4^{d-1}}{d}\right] = \frac{1}{2d} .$$

□

We now have everything we need to provide reasonably tight upper bounds for the $\sigma$ coefficients.

**Proposition 6 (Bounding the $\sigma$ coefficients).** *Let $d \geqslant 2$. Then the following holds:*

$$|\sigma_0| \leqslant \frac{3d}{4}\,, \quad |\sigma_1| \leqslant \frac{1}{2}\,, \quad |\sigma_2| \leqslant 2\mathrm{e}^{\frac{1}{2\nu^2}}\,, \quad \textit{and} \quad |\sigma_3| \leqslant \frac{2\mathrm{e}^{\frac{1}{2\nu^2}}}{d} .$$

*Proof.* From Lemma 1 and the definition of $\sigma_0$, we have

$$|\sigma_0| = |(d-1)\alpha_2 + \alpha_1| \leqslant \frac{d-1}{4} + \frac{1}{2} = \frac{d+3}{4} .$$

We deduce the first result since $d \geqslant 2$. Next, since $\sigma_1 = -\alpha_1$, we obtain $|\sigma_1| \leqslant 1/2$ directly from Lemma 1. Regarding the last two coefficients, recall that Proposition 3 guarantees that their common denominator $\alpha_1 - \alpha_2$ is lower bounded by $\mathrm{e}^{\frac{-1}{2\nu^2}}/4$. Since

$$(d-2)\alpha_0\alpha_2 - (d-1)\alpha_1^2 + \alpha_0\alpha_1 = c_d + \alpha_1^2 - \alpha_0\alpha_2\,,$$

we can write $\sigma_2 = (c_d + \alpha_1^2 - \alpha_0\alpha_2)/(\alpha_1 - \alpha_2)$ and deduce that

$$|\sigma_2| \leqslant \frac{1/4 + 1/(2d)}{\mathrm{e}^{\frac{-1}{2\nu^2}}/4} \leqslant 2\mathrm{e}^{\frac{1}{2\nu^2}}\,,$$

since, according to Eq. (7), $c_d \leqslant 1/4$ and $\alpha_1^2 - \alpha_0\alpha_2 \leqslant 1/(2d)$ according to Lemma 2. Finally, we write

$$|\sigma_3| = \left|\frac{\alpha_1^2 - \alpha_0\alpha_2}{\alpha_1 - \alpha_2}\right| \leqslant \frac{1/(2d)}{\mathrm{e}^{\frac{-1}{2\nu^2}}/4} = \frac{2\mathrm{e}^{\frac{1}{2\nu^2}}}{d} .$$

□

The bounds obtained in Proposition 6 immediately translate into a control of the Frobenius norm of $\Sigma^{-1}$, which in turn yields a control over the operator norm of $\Sigma^{-1}$, as promised.

**Corollary 1 (Control of $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}$).** *Let $d \geqslant 2$. Then $\left\|\Sigma^{-1}\right\|_{\mathrm{op}} \leqslant 8d\mathrm{e}^{\frac{1}{\nu^2}}$.*

*Proof.* Using Proposition 6, we write

$$\begin{aligned}
\left\|\Sigma^{-1}\right\|_{\mathrm{F}}^2 &= \frac{1}{c_d^2}\left[\sigma_0^2 + 2d\sigma_1^2 + d\sigma_2^2 + (d^2-d)\sigma_3^2\right] \\
&\leqslant 16\mathrm{e}^{\frac{1}{\nu^2}}\left[\frac{9d^2}{16} + \frac{2d}{4} + 4d\mathrm{e}^{\frac{1}{\nu^2}} + 4\mathrm{e}^{\frac{1}{\nu^2}}\right] \\
&\leqslant 61 d^2 \mathrm{e}^{\frac{2}{\nu^2}} \, ,
\end{aligned}$$

where we used $d \geqslant 2$ in the last display. Since the operator norm is upper bounded by the Frobenius norm, we conclude observing that $\sqrt{61} \leqslant 8$. □

**Remark 1.** The bound on $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}$ is essentially tight with respect to the dependency in $d$, as can be seen in Figure 4.

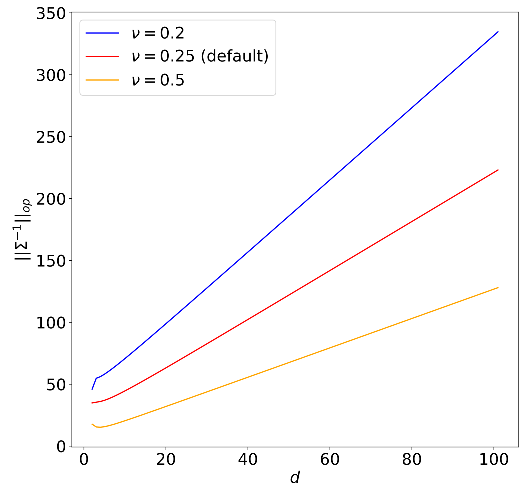


Figure 4: Evolution of $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}$ as a function of $d$ for various values of the bandwidth parameter. The linear dependency in $d$ is striking.

## 2 Study of $\hat{\Gamma}_n$

We now turn to the study of $\hat{\Gamma}_n$. We start by computing the limiting expression. Recall that we defined $\hat{\Gamma}_n = \frac{1}{n} Z^\top W y$, where $y \in \mathbb{R}^{d+1}$ is the random vector defined coordinate-wise by $y_i = f(x_i)$. From the definition of $\hat{\Gamma}_n$, it is straightforward that

$$\hat{\Gamma}_n = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \pi_i f(x_i) \\ \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,1} f(x_i) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n \pi_i z_{i,d} f(x_i) \end{pmatrix} \in \mathbb{R}^{d+1} \, .$$

As a consequence, if we define $\Gamma^f := \mathbb{E}[\hat{\Gamma}_n]$, it holds that

$$\Gamma^f = \begin{pmatrix} \mathbb{E}\left[\pi f(x)\right] \\ \mathbb{E}\left[\pi z_1 f(x)\right] \\ \vdots \\ \mathbb{E}\left[\pi z_d f(x)\right] \end{pmatrix} \, . \tag{16}$$

We specialize Eq. (16) to shape detectors in Section 2.1 and linear models in Section 2.2. The concentration of $\hat{\Gamma}_n$ around $\Gamma$ is obtained in Section 2.3.

### 2.1 Shape detectors

Recall that we defined

$$\forall x \in [0,1]^D, \quad f(x) = \prod_{u \in \mathcal{S}} \mathbf{1}_{x_u > \tau} \, , \tag{17}$$

with $\mathcal{S} = \{u_1, \ldots, u_q\}$ a fixed set of pixels indices and $\tau \in (0,1)$ a threshold. As in the paper, let us define $E = \{j \text{ s.t. } J_j \cap \mathcal{S} \neq \varnothing\}$ denote the set of superpixels intersecting the shape, and

$$E_+ = \{j \in E \text{ s.t. } \overline{\xi}_j > \tau\} \quad \text{and} \quad E_- = \{j \in E \text{ s.t. } \overline{\xi}_j \leqslant \tau\} \, .$$

We also defined

$$\mathcal{S}_+ = \{u \in \mathcal{S} \text{ s.t. } \xi_u > \tau\} \quad \text{and} \quad \mathcal{S}_- = \{u \in \mathcal{S} \text{ s.t. } \xi_u \leqslant \tau\} .$$

In the main paper, we made the following simplifying assumption:

$$\forall j \in E_+, \quad J_j \cap \mathcal{S}_- = \varnothing . \tag{18}$$

This is not the case here. Unfortunately, without this assumption, the expression of $\Gamma^f$ is slightly more complicated and we need to generalize the definition of the $\alpha$ coefficients.

**Definition 1 (Generalized $\alpha$ coefficients).** For any $p, q$ such that $p + q \leqslant d$, we define

$$\alpha_{p,q} := \mathbb{E}\left[\pi z_1 \cdots z_p \cdot (1 - z_{p+1}) \cdots (1 - z_{p+q})\right] . \tag{19}$$

We notice that, for any $1 \leqslant p \leqslant d$, $\alpha_{p,0} = \alpha_p$. As it is the case with $\alpha$ coefficients, the generalized $\alpha$ coefficients can be computed in closed-form:

**Proposition 7 (Computation of the generalized $\alpha$ coefficients).** *Let $p, q$ such that $p+q \leqslant d$. Then*

$$\alpha_{p,q} = \frac{1}{2^d} \sum_{s=0}^{d} \binom{d-p-q}{s-q} \psi\left(\frac{s}{d}\right) .$$

*Proof.* We follow the proof of Proposition 1.

$$\begin{aligned}
\alpha_{p,q} &= \mathbb{E}\left[\pi z_1 \cdots z_p \cdot (1 - z_{p+1}) \cdots (1 - z_{p+q})\right] \\
&= \sum_{s=0}^{d} \mathbb{E}_s\left[\pi z_1 \cdots z_p \cdot (1 - z_{p+1}) \cdots (1 - z_{p+q})\right] \cdot \mathbb{P}\left(S = s\right) \\
&= \frac{1}{2^d} \sum_{s=0}^{d} \binom{d}{s} \psi\left(\frac{s}{d}\right) \mathbb{P}_s\left(z_1 = \cdots = z_p = 1, z_{p+1} = \cdots = z_{p+q} = 0\right) \\
&= \frac{1}{2^d} \sum_{s=0}^{d} \binom{d}{s} \psi\left(\frac{s}{d}\right) \binom{d-p-q}{s-q} \binom{d}{s} \qquad \text{(Lemma 4)} \\
\alpha_{p,q} &= \frac{1}{2^d} \sum_{s=0}^{d} \binom{d-p-q}{s-q} \psi\left(\frac{s}{d}\right) .
\end{aligned}$$

□

Notice that the expression of $\alpha_{p,q}$ coincide with that of $\alpha_p$ when $q = 0$. We can now give the expression of $\Gamma^f$ for an elementary shape detector in the general case.

**Proposition 8 (Computation of $\Gamma^f$, elementary shape detector).** *Assume that $f$ is written as in Eq. (17). Assume that for any $j \in E_-$, $J_j \cap \mathcal{S}_- = \varnothing$ (otherwise $\Gamma^f = 0$). Let $p := |E_-|$ and $q := |\{j \in E_+, J_j \cap \mathcal{S}_- \neq \varnothing\}|$. Then $\mathbb{E}\left[\pi f(x)\right] = \alpha_{p,q}$ and*

$$\mathbb{E}\left[\pi z_j f(x)\right] = \begin{cases} 0 & \text{if } j \in \{j \in E_+ \text{ s.t. } J_j \cap \mathcal{S}_- \neq \varnothing\} , \\ \alpha_{p,q} & \text{if } j \in E_- , \\ \alpha_{p+1,q} & \text{otherwise.} \end{cases}$$

Taking $q = 0$ (a consequence of Eq. (18)) in Proposition 8 directly yields $\mathbb{E}\left[\pi f(x)\right] = \alpha_p$ and

$$\mathbb{E}\left[\pi z_j f(x)\right] = \begin{cases} \alpha_p & \text{if } j \in E_- , \\ \alpha_{p+1} & \text{otherwise.} \end{cases}$$

*Proof.* We notice that, for any $u \in J_j$,

$$x_u = z_j \xi_u + (1 - z_j)\overline{\xi}_u .$$

There are four cases to consider when deciding whether $x_u > \tau$ or not:

- $\xi_u > \tau$ and $\overline{\xi}_u > \tau$, that is, $j \in E_+$ and $u \in J_j \cap \mathcal{S}_+$. Then $x_u > \tau$ a.s.;
- $\xi_u \leqslant \tau$ and $\overline{\xi}_u > \tau$, that is, $j \in E_+$ and $u \in J_j \cap \mathcal{S}_-$. Then $x_u > \tau$ if, and only if, $z_j = 0$;
- $\xi_u > \tau$ and $\overline{\xi}_u \leqslant \tau$, that is, $j \in E_-$ and $u \in J_j \cap \mathcal{S}_+$. Then $x_u > \tau$ if, and only if, $z_j = 1$;

- $\xi_u \leqslant \tau$ and $\overline{\xi}_u \leqslant \tau$, that is, $j \in E_-$ and $u \in J_j \cap \mathcal{S}_-$. Then $x_u \leqslant \tau$ a.s., but this last case cannot happen since we assume that for any $j \in E_-$, $J_j \cap \mathcal{S}_- = \varnothing$.

This case separation allows us to rewrite $f(x)$ as

$$\begin{aligned} f(x) &= \prod_{u \in \mathcal{S}} \mathbf{1}_{x_u > \tau} && \text{(Eq. (17))} \\ &= \prod_{j \in E_+} \prod_{u \in J_j \cap \mathcal{S}_-} (1 - z_j) \cdot \prod_{j \in E_-} \prod_{u \in J_j \cap \mathcal{S}_+} z_j \end{aligned}$$

Since we assumed that for any $j \in E_-$, $J_j \cap \mathcal{S}_- = \varnothing$, then for any $j \in E_-$, $J_j \cap \mathcal{S}_+ \neq \varnothing$. Thus the rightmost inner products are never empty, and since $z_j \in \{0, 1\}$ a.s., we deduce that there are $p$ terms in the rightmost product. By definition of $q$, and again since $1 - z_j \in \{0, 1\}$ a.s., there are $q$ terms in the leftmost product. By definition of $E_+$ and $E_-$, these products do not have any common terms. We deduce that $\mathbb{E}[\pi f(x)] = \alpha_{p,q}$ by definition of the generalized $\alpha$ coefficients.

When computing $\mathbb{E}[\pi z_j f(x)]$, there are several possibilities. First, if $j \in \{j \in E_+ \text{ s.t. } J_j \cap \mathcal{S}_- \neq \varnothing\}$, since $z_j(1 - z_j) = 0$ a.s., we deduce that $\mathbb{E}[\pi z_j f(x)] = 0$. Second, if $j \in E_-$, since $z_j^2 = z_j$, we recover $\mathbb{E}[\pi z_j f(x)] = \mathbb{E}[\pi f(x)] = \alpha_{p,q}$. Finally, if $j$ does not belong to one of these sets, then the rightmost product gains one additional term and we obtain $\alpha_{p+1,q}$. ☐

## 2.2 Linear model

In this section, we compute $\Gamma^f$ for a linear $f$. As in the paper, we define

$$f(x) = \sum_{u=1}^{D} \lambda_u x_u \,, \tag{20}$$

with $\lambda_1, \ldots, \lambda_D \in \mathbb{R}$ arbitrary coefficients. By linearity, we just have to look into the case $f : x \mapsto x_u$ where $u \in \{1, \ldots, D\}$ is a fixed pixel index.

**Proposition 9 (Computation of $\Gamma^f$, linear case).** *Assume that $f$ is defined as in Eq. (20) and $u \in J_j$. Then*

$$\mathbb{E}[\pi x_u] = \alpha_1(\xi_u - \overline{\xi}_u) + \alpha_0 \overline{\xi}_u \,,$$
$$\mathbb{E}[\pi z_j x_u] = \alpha_1(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u \,,$$

*and, for any $k \neq j$,*

$$\mathbb{E}[\pi z_k x_u] = \alpha_2(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u \,.$$

*Proof.* As in the proof of Proposition 8, we notice that

$$x_u = z_j \xi_u + (1 - z_j)\overline{\xi}_u \,.$$

Then we write

$$\begin{aligned} \mathbb{E}[\pi x_u] &= \mathbb{E}\left[\pi(z_j \xi_u + (1 - z_j)\overline{\xi}_u)\right] \\ &= \mathbb{E}\left[\pi z_j(\xi_u - \overline{\xi}_u) + \pi \overline{\xi}_u\right] \\ \mathbb{E}[\pi x_u] &= \alpha_1(\xi_u - \overline{\xi}_u) + \alpha_0 \overline{\xi}_u \,, \end{aligned}$$

where we used the definition of the $\alpha$ coefficients. Now let us compute $\mathbb{E}[\pi z_j f(x)]$:

$$\begin{aligned} \mathbb{E}[\pi z_j x_u] &= \mathbb{E}\left[\pi z_j(z_j \xi_u + (1 - z_j)\overline{\xi}_u)\right] \\ &= \mathbb{E}\left[\pi z_j((\xi_u - \overline{\xi}_u) z_j + \overline{\xi}_u)\right] && (z_j \in \{0, 1\} \text{ a.s.}) \\ \mathbb{E}[\pi z_j x_u] &= \alpha_1(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u \,. \end{aligned}$$

And finally, for any $k \neq j$,

$$\begin{aligned} \mathbb{E}[\pi z_k x_u] &= \mathbb{E}\left[\pi z_k((\xi_u - \overline{\xi}_u) z_j + \overline{\xi}_u)\right] \\ &= \alpha_2(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u \,. \end{aligned}$$

☐

### 2.3 Concentration of $\hat{\Gamma}_n$

We now show that $\hat{\Gamma}_n$ is concentrated around $\Gamma^f$. Since the expression of $\hat{\Gamma}_n$ is the same than in the tabular case, and we assume that $f$ is bounded on the support of $x$, the same reasoning as in the proof of Proposition 24 in Garreau and von Luxburg [2020] can be applied.

**Proposition 10 (Concentration of $\hat{\Gamma}_n$).** *Assume that $f$ is bounded by $M > 0$ on Supp$(x)$. Then, for any $t > 0$, it holds that*

$$\mathbb{P}\left(\|\hat{\Gamma}_n - \Gamma^f\| \geqslant t\right) \leqslant 4d\exp\left(\frac{-nt^2}{32Md^2}\right).$$

*Proof.* Since $f$ is bounded by $M$ on Supp$(x)$, it holds that $|f(x)| \leqslant M$ almost surely. We can then proceed as in the proof of Proposition 24 in Garreau and von Luxburg [2020]. □

## 3 The study of $\beta^f$

### 3.1 Concentration of $\hat{\beta}_n$

In this section we show the concentration of $\hat{\beta}_n$ (Theorem 1 in the paper). The proof scheme follows closely that of Garreau and von Luxburg [2020].

**Theorem 1 (Concentration of $\hat{\beta}_n$).** *Assume that $f$ is bounded by a constant $M$ on the unit cube $[0,1]^D$. Let $\epsilon > 0$ and $\eta \in (0,1)$. Let $d$ be the number of superpixels used by LIME. Then, there exists $\beta^f \in \mathbb{R}^{d+1}$ such that, for every*

$$n \geqslant \left\lceil \max\left(2^{15}d^4\mathrm{e}^{\frac{2}{\nu^2}}, \frac{2^{21}d^7\max(M,M^2)\mathrm{e}^{\frac{4}{\nu^2}}}{\epsilon^2}\right)\log\frac{8d}{\eta}\right\rceil,$$

*we have $\mathbb{P}(\|\hat{\beta}_n - \beta^f\| \geqslant \epsilon) \leqslant \eta$.*

*Proof.* As in Garreau and von Luxburg [2020], the key idea of the proof is to notice that

$$\|\hat{\beta}_n - \beta^f\| \leqslant 2\left\|\Sigma^{-1}\right\|_{\mathrm{op}}\|\hat{\Gamma} - \Gamma^f\| + 2\left\|\Sigma^{-1}\right\|_{\mathrm{op}}^2\left\|\Gamma^f\right\|\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \tag{21}$$

provided that (i) $\|\Sigma^{-1}(\hat{\Sigma} - \Sigma)\|_{\mathrm{op}} \leqslant 0.32$ (this is Lemma 27 in Garreau and von Luxburg [2020]. We are going to build an event of probability at least $1-\eta$ such that $\hat{\Sigma}_n$ is close to $\Sigma$ and $\hat{\Gamma}_n$ is close from $\Gamma^f$. The deterministic bound obtained on $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}$ together with the boundedness of $f$ will allow us to show that (ii) $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}\|\hat{\Gamma} - \Gamma^f\| \leqslant \epsilon/4$ and (iii) $\left\|\Sigma^{-1}\right\|_{\mathrm{op}}^2\left\|\Gamma^f\right\|\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} \leqslant \epsilon/4$.

We first show (i). Let us set $n_1 := \left\lceil 2^{15}d^4\mathrm{e}^{\frac{2}{\nu^2}}\log\frac{8d}{\eta}\right\rceil$ and $t_1 := \frac{1}{25d\mathrm{e}^{\frac{1}{\nu^2}}}$. According to Proposition 5, for any $n \geqslant n_1$,

$$\mathbb{P}\left(\left\|\hat{\Sigma}_n - \Sigma\right\|_{\mathrm{op}} \geqslant t_1\right) \leqslant 4d\cdot\exp\left(\frac{-n_1t_1^2}{32d^2}\right) \leqslant \frac{\eta}{2}.$$

Moreover, we know that $\left\|\Sigma^{-1}\right\|_{\mathrm{op}} \leqslant 8d\mathrm{e}^{\frac{1}{\nu^2}}$ (Corollary 1). Since the operator norm is sub-multiplicative, with probability greater than $1-\eta/2$, we have

$$\left\|\Sigma^{-1}(\hat{\Sigma}_n - \Sigma)\right\|_{\mathrm{op}} \leqslant \left\|\Sigma^{-1}\right\|_{\mathrm{op}}\cdot\left\|\hat{\Sigma}_n - \Sigma\right\|_{\mathrm{op}} \leqslant 8d\mathrm{e}^{\frac{1}{\nu^2}}\cdot t_1 = 0.32.$$

Now let us show (ii). Let us define $n_2 := \left\lceil\frac{2^{15}Md^4\mathrm{e}^{\frac{2}{\nu^2}}}{\epsilon^2}\log\frac{8d}{\eta}\right\rceil$ and $t_2 := \frac{\epsilon}{32d\mathrm{e}^{\frac{1}{\nu^2}}}$. According to Proposition 10, for any $n \geqslant n_2$, we have

$$\mathbb{P}\left(\left\|\hat{\Gamma}_n - \Gamma\right\| \geqslant t_2\right) \leqslant 4d\cdot\exp\left(\frac{-n_2t_2^2}{32Md^2}\right) \leqslant \frac{\eta}{2}.$$

Recall that $\left\|\Sigma^{-1}\right\|_{\mathrm{op}} \leqslant 8d\mathrm{e}^{\frac{1}{\nu^2}}$ (Corollary 1): with probability higher than $1-\eta/2$,

$$\left\|\Sigma^{-1}\right\|_{\mathrm{op}}\cdot\left\|\hat{\Gamma}_n - \Gamma^f\right\| \leqslant 8d\mathrm{e}^{\frac{1}{\nu^2}}\cdot t_2 = \frac{\epsilon}{4}.$$

Finally let us show (iii). Let us define $n_3 := \left\lceil\frac{2^{21}d^7M^2\mathrm{e}^{\frac{4}{\nu^2}}}{\epsilon^2}\log\frac{8d}{\eta}\right\rceil$ and $t_3 := \frac{\epsilon}{2^8Md^{5/2}\mathrm{e}^{\frac{2}{\nu^2}}}$. According to Proposition 5, for any $n \geqslant n_3$, we have

$$\mathbb{P}\left(\left\|\hat{\Sigma}_n - \Sigma\right\|_{\mathrm{op}} \geqslant t_3\right) \leqslant 4d\cdot\exp\left(\frac{-n_3t_3^2}{32d^2}\right) \leqslant \frac{\eta}{2}.$$

Since $f$ is bounded by $M$, it is straightforward to show that $\left\|\hat{\Gamma}^f\right\| \leqslant M \cdot d^{1/2}$. Moreover, recall that $\left\|\Sigma^{-1}\right\|_{\text{op}}^2 \leqslant 64d^2\mathrm{e}^{\frac{2}{\nu^2}}$. We deduce that, with probability at least $\eta/2$,

$$\left\|\Sigma^{-1}\right\|_{\text{op}}^2 \cdot \left\|\Gamma^f\right\| \cdot \left\|\hat{\Sigma}_n - \Sigma\right\|_{\text{op}} \leqslant 64d^2\mathrm{e}^{\frac{2}{\nu^2}} \cdot Md^{1/2} \cdot t_3 = \frac{\epsilon}{4}\,.$$

Finally, we notice that both $n_2$ and $n_3$ are smaller than

$$n_4 := \left\lceil \frac{2^{21}d^7\max(M, M^2)\mathrm{e}^{\frac{4}{\nu^2}}}{\epsilon^2} \log\frac{8d}{\eta}\right\rceil\,.$$

Thus (ii) and (ii) simultaneously happen on an event of probability greater than $\eta/2$ when $n$ is larger than $n_4$. We conclude by a union bound argument. □

**Remark 2.** In view of Remark 1, it seems difficult to improve much the rate of convergence given by Theorem 1 with the current proof technology. Indeed, a careful inspection of the proof reveals that, starting from Eq. (21), the control of $\left\|\Sigma^{-1}\right\|_{\text{op}}$ is key. Since the dependency in $d$ seems tight, there is not much hope for improvement.

### 3.2 General expression of $\beta^f$

We are now able to recover Proposition 2 of the paper: the expression of $\beta^f$ is obtained simply by multiplying Eq. (3) and (16). We also give the value of the intercept ($\beta_0$ with our notation), which is omitted in the paper for simplicity's sake.

**Corollary 2 (Computation of $\beta^f$).** *Under the assumptions of Theorem 1.*

$$\beta_0^f = c_d^{-1}\left\{\sigma_0\mathbb{E}\left[\pi f(x)\right] + \sigma_1\sum_{j=1}^d \mathbb{E}\left[\pi z_j f(x)\right]\right\}, \tag{22}$$

*and, for any $1 \leqslant j \leqslant d$,*

$$\beta_j^f = c_d^{-1}\left\{\sigma_1\mathbb{E}\left[\pi f(x)\right] + \sigma_2\mathbb{E}\left[\pi z_j f(x)\right] + \sigma_3\sum_{\substack{k=1\\k\neq j}}^d \mathbb{E}\left[\pi z_k f(x)\right]\right\}. \tag{23}$$

### 3.3 Shape detectors

We now specialize Corollary 2 to the case of elementary shape detectors.

**Proposition 11 (Expression of $\beta^f$, shape detector).** *Let $f$ be written as in Eq. (17). Assume that for any $j \in E_-$, $J_j \cap \mathcal{S}_- = \varnothing$ (otherwise $\beta^f = 0$). Let $p$ and $q$ as before. Then*

$$\beta_0^f = c_d^{-1}\left\{\sigma_0\alpha_{p,q} + p\sigma_1\alpha_{p,q} + (d-p-q)\alpha_{p+1,q}\right\}\,,$$

*for any $j \in E_-$,*

$$\beta_j^f = c_d^{-1}\left\{\sigma_1\alpha_{p,q} + \sigma_2\alpha_{p,q} + (p-1)\sigma_2\alpha_{p,q} + (d-p-q)\sigma_3\alpha_{p+1,q}\right\}\,,$$

*for any $j \in E_+$ such that $J_j \cap \mathcal{S}_- \neq \varnothing$,*

$$\beta_j^f = c_d^{-1}\left\{\sigma_1\alpha_{p,q} + p\sigma_3\alpha_{p,q} + (d-p-q)\alpha_{p+1,q}\right\}\,,$$

*and*

$$\beta_j^f = c_d^{-1}\left\{\sigma_1\alpha_{p,q} + \sigma_2\alpha_{p+1,q} + p\sigma_3\alpha_{p,q} + (d-p-q-1)\sigma_3\alpha_{p+1,q}\right\}$$

*otherwise.*

*Proof.* Straightforward from Corollary 2 and Proposition 8. □

Note that taking $q = 0$ in Proposition 11 yields Proposition 3 of the paper.

### 3.4 Linear models

We deduce from Proposition 9 the expression of $\beta^f$ for linear models. Let us define $M_j$ the binary mask associated to superpixel $J_j$ and let $\circ$ be the termwise product.

**Proposition 12 (Computation of $\beta^f$, linear case).** *Assume that $f$ is defined as in Eq.* (20)*. Then*

$$\beta_0^f = \sum_{u=1}^{D} \lambda_u \overline{\xi}_u = f(\overline{\xi}) \, ,$$

*and, for any* $1 \leqslant j \leqslant d$*,*

$$\beta_j^f = \sum_{u \in J_j} \lambda_u (\xi_u - \overline{\xi}_u) = f(M_j \circ (\xi - \overline{\xi})) \, .$$

It is interesting to compute prediction of the surrogate model at $\xi$:

$$\beta_0^f + \beta_1^f + \cdots + \beta_d^f = f(\overline{\xi}) + f(M_1 \circ (\xi - \overline{\xi})) + \cdots + f(M_d \circ (\xi - \overline{\xi})) = f(\xi) \, .$$

Thus in the case of linear models, the limit explanation is faithful.

*Proof.* By linearity, we can start by computing $\beta^f$ for the function $x \mapsto x_u$. Assume that $j \in \{1, \ldots, d\}$ is such that $u \in J_j$. According to Corollary 2 and Proposition 9,

$$\begin{aligned}
\beta_0^f &= \frac{1}{c_d} \left\{ \sigma_0 \mathbb{E}\left[\pi f(x)\right] + \sigma_1 \sum_{j=1}^{d} \mathbb{E}\left[\pi z_j f(x)\right] \right\} \\
&= \frac{1}{c_d} \left\{ \sigma_0(\alpha_1(\xi_u - \overline{\xi}_u) + \alpha_0 \overline{\xi}_u) + \sigma_1(\alpha_1(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u) + (d-1)\sigma_1(\alpha_2(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u) \right\} \\
&= \frac{1}{c_d} \left\{ (\sigma_0\alpha_1 + \sigma_1\alpha_1 + (d-1)\sigma_1\alpha_2)(\xi_u - \overline{\xi}_u) + (\sigma_0\alpha_0 + d\sigma_1\alpha_1)\overline{\xi}_u \right\} \\
\beta_0^f &= \overline{\xi}_u \, ,
\end{aligned}$$

where we used Eqs. (8) and (12) in the last display.

$$\begin{aligned}
\beta_j^f &= \frac{1}{c_d} \left\{ \sigma_1 \mathbb{E}\left[\pi f(x)\right] + \sigma_2 \mathbb{E}\left[\pi z_j f(x)\right] + \sigma_3 \sum_{\substack{k=1 \\ k \neq j}}^{d} \mathbb{E}\left[\pi z_k f(x)\right] \right\} \\
&= \frac{1}{c_d} \left\{ \sigma_1(\alpha_1(\xi_u - \overline{\xi}_u) + \alpha_0 \overline{\xi}_u) + \sigma_2(\alpha_1(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u) + (d-1)\sigma_3(\alpha_2(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u) \right\} \\
&= \frac{1}{c_d} \left\{ (\sigma_1\alpha_1 + \sigma_2\alpha_1 + (d-1)\sigma_3\alpha_2)(\xi_u - \overline{\xi}_u) + (\sigma_1\alpha_0 + \sigma_2\alpha_1 + (d-1)\sigma_3\alpha_1)\overline{\xi}_u \right\} \\
\beta_j^f &= \xi_u - \overline{\xi}_u \, ,
\end{aligned}$$

where we used Eqs. (9) and (11) in the last display. Finally, let $k \neq j$:

$$\begin{aligned}
\beta_k^f &= \frac{1}{c_d} \left\{ \sigma_1 \mathbb{E}\left[\pi f(x)\right] + \sigma_2 \mathbb{E}\left[\pi z_k f(x)\right] + \sigma_3 \sum_{\substack{k'=1 \\ k' \neq j,k}}^{d} \mathbb{E}\left[\pi z_{k'} f(x)\right] \right\} \\
&= \frac{1}{c_d} \Big\{ \sigma_1(\alpha_1(\xi_u - \overline{\xi}_u) + \alpha_0 \overline{\xi}_u) + \sigma_2(\alpha_2(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u) + \sigma_3(\alpha_1(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u) \\
&\qquad + (d-2)\sigma_3(\alpha_2(\xi_u - \overline{\xi}_u) + \alpha_1 \overline{\xi}_u) \Big\} \\
&= \frac{1}{c_d} \left\{ (\sigma_1\alpha_1 + \sigma_2\alpha_2 + \sigma_3\alpha_1 + (d-2)\sigma_3\alpha_2)(\xi_u - \overline{\xi}_u) + (\sigma_1\alpha_0 + \sigma_2\alpha_1 + (d-1)\sigma_3\alpha_1)\overline{\xi}_u \right\} \\
\beta_k^f &= 0 \, ,
\end{aligned}$$

where we used Eqs. (10) and (11) in the last display. We deduce the result by linearity. □

# 4 Technical results

### 4.1 Probability computations

In this section we collect all elementary probability computations necessary for the computation of the $\alpha$ coefficients and the generalized $\alpha$ coefficients.

**Lemma 3 (Activated only).** *Let $p \geqslant 0$ be an integer. Then*

$$\mathbb{P}_s(z_1 = 1, \dots, z_p = 1) = \frac{(d-p)!}{d!} \cdot \frac{(d-s)!}{(d-s-p)!} \,.$$

*Proof.* Conditionally to $S = s$, the choice of $S$ is uniform among all subsets of $\{1, \dots, d\}$. Therefore we recover the proof of Lemma 4 in Mardaoui and Garreau [2021]. □

The following lemma is a slight generalization, which coincides when $q = 0$.

**Lemma 4 (Activated and deactivated).** *Let $p, q$ be integers. Then*

$$\mathbb{P}_s(z_1 = \cdots = z_p = 1, z_{p+1} = \cdots = z_{p+q} = 0) = \binom{d-p-q}{s-q}\binom{d}{s}^{-1} \,.$$

*Proof.* Conditionally to $S = s$, the deletions are uniformly distributed. Therefore, the total number of cases is $\binom{d}{s}$. Now, the favorable cases correspond to superpixels $p+1, \dots, p+q$ deleted: these are $q$ fixed deletions. We also need to have superpixels $1, \dots, p$ activated, these are $p$ indices that are not available to deletions. In total, we need to place $s-q$ deletions among $d-p-q$ possibilities. We deduce the result. □

## 4.2 Algebraic identities

In this section we collect some identities used throughout the proofs.

**Proposition 13 (Four letter identity).** *Let $A$, $B$, $C$, and $D$ be four finite sequences of real numbers. Then it holds that*

$$\sum_j A_j C_j \cdot \sum_j B_j D_j - \sum_j A_j B_j \cdot \sum C_j D_j = \sum_{j<k} (A_j D_k - A_k D_j)(C_j B_k - C_k B_j) \,.$$

*Proof.* See the proof of Exercise 3.7 in Steele [2004]. □

**Proposition 14 (A combinatorial identity).** *Let $d \geqslant 1$ be an integer. Then*

$$V_d := \sum_{j<k} \binom{d}{j}\binom{d}{k}(j-k)^2 = d \cdot 4^{d-1} \,.$$

*Proof.* We first notice that

$$\begin{aligned} V_d &= \frac{1}{2}\sum_{j,k} \binom{d}{j}\binom{d}{k}(j-k)^2 && \text{(by symmetry)} \\ &= \sum_{j,k} \binom{d}{j}\binom{d}{k}k^2 - \sum_{j,k}\binom{d}{j}\binom{d}{k} jk && \text{(developing the square)} \\ &= \sum_j \binom{d}{j} \sum_k \binom{d}{k} k^2 - \left(\sum_j \binom{d}{j} j\right)^2 \,. \end{aligned}$$

It is straightforward to show that

$$\sum_j \binom{d}{j} = 2^d \,, \sum_j \binom{d}{j} j = d \cdot 2^{d-1} \,, \text{ and } \sum_j \binom{d}{j} j^2 = d(d+1) \cdot 2^{d-2} \,.$$

We deduce that

$$c_d = 2^d \cdot d(d+1) \cdot 2^{d-2} - d^2 \cdot 2^{2d-2} = d \cdot 4^{d-1} \,.$$

□

# 5 Additional results

In this section, we present additional qualitative results on the three pre-trained models used in the paper: MobileNetV2 [Sandler et al., 2018], DenseNet121 [Huang et al., 2017], and InceptionV3 [Szegedy et al., 2016].

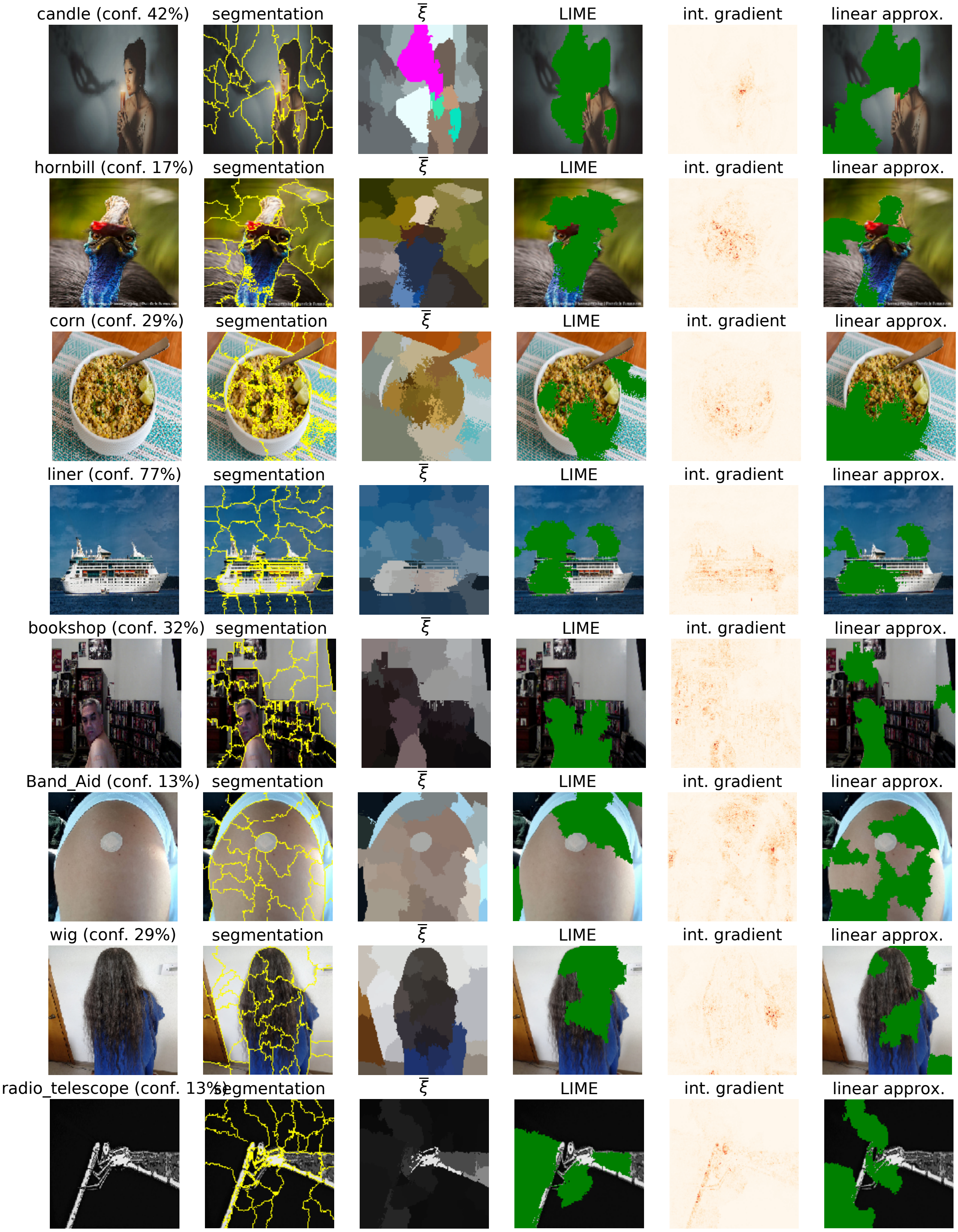


Figure 5: Empirical explanations, integrated gradient, and approximated explanations for images from the ILSVRC2017 dataset. The model explained is the likelihood function associated to the top class given by MobileNetV2.


Figure 6: Empirical explanations, integrated gradient, and approximated explanations for images from the ILSVRC2017 dataset. The model explained is the likelihood function associated to the top class given by DenseNet121.

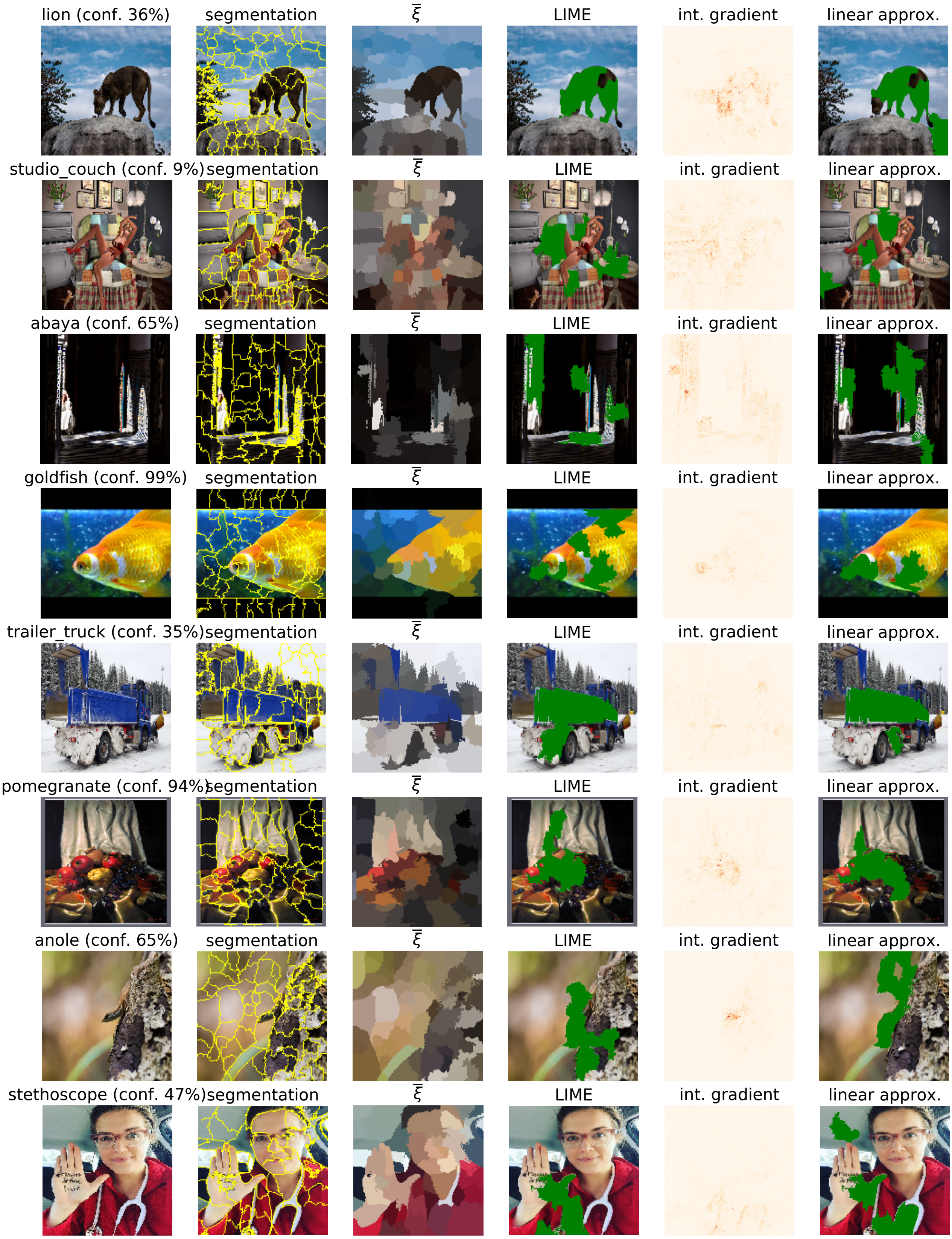


Figure 7: Empirical explanations, integrated gradient, and approximated explanations for images from the ILSVRC2017 dataset. The model explained is the likelihood function associated to the top class given by InceptionV3.

# References

D. Garreau and U. von Luxburg. Looking Deeper into Tabular LIME. *arXiv preprint arXiv:2008.11092*, 2020.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

D. Mardaoui and D. Garreau. An Analysis of LIME for Text Data. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

J. M. Steele. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.