Cluster Stability for Finite Samples Supplementary Material

Ohad Shamir[†] **and Naftali Tishby**[†‡]
† School of Computer Science and Engineering
‡ Interdisciplinary Center for Neural Computation
The Hebrew University
Jerusalem 91904, Israel
{ohadsh,tishby}@cs.huji.ac.il

Abstract

This technical report contains the proofs of Lemmas 1 and 2 in the paper 'Cluster Stability for Finite Samples'.

1 Proof of Lemma 1

Proof. The proof idea is essentially identical for all values of k. We have that $d_{\mathcal{D}}(A_k(S_1), A_k(S_2))$ is governed by the the probability mass of $\mathcal{D}$ which switches between clusters in $A_k(S_1)$ and $A_k(S_2)$, in expectation over S_1 and S_2. For reasonably large samples, all this probability mass is tightly concentrated in small border regions between the clusters, and is governed by small fluctuations in the border positions. For all k, these fluctuations become smaller as the sample size m increases. The important point is that the location of the border points are different for different choices of k. For the 'right' model, the borders lie in areas of very low probability density, and as a result the probability mass of $\mathcal{D}$ which switches between clusters is relatively small in expectation. In contrast, for the 'wrong' models, some of the border points lie in areas of higher density, so the probability mass of $\mathcal{D}$ which switches between clusters is relatively much higher. From this, we get that $stab(A_k, \mathcal{D}, m)$ is relatively smaller for the 'right' value of k, compared to the other values.

We will consider the case $k = 2$ in some detail, and then go over the other two cases more quickly. To simplify the analysis, the proof involves some approximations, with approximation errors which are asymptotically negligible as $m \to \infty$, or that are arbitrarily small if μ is large enough. Approximations of the first type form the $o(1)$ term in the lemma, while approximations of the second type can be absorbed into the derived (non-tight) bounds. We will use the formulation $\mathcal{N}(\mu, \sigma^2)$ to denote a normally distributed real random variable, with expectation μ and variance σ^2. Also, we will make frequent use of the following basic facts: If a_1, a_2 are independent random variables such that $a_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $b \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then the distribution of $a_1 + a_2$ is $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, and the expected value of $|a_1|$ is $\sqrt{2/\pi}\sigma_1$.

For $k = 2$, let α_1 and α_2 be random variables (over the draw of a sample of size m from $\mathcal{D}$), representing the centroids in $\mathbb{R}$ returned by the algorithm, such that $\alpha_1 \leq \alpha_2$ (see figure 1). If the Gaussians are well separated, we can assume that they are approximately independent: the value of α_1 is equal to the sample mean derived from the region of the larger Gaussian, while α_2 is equal to the sample mean derived from the mixture of the two smaller Gaussians. The distribution of a sample mean of a unit variance Gaussian is also Gaussian, with variance $1/n$ where n is the sample size on which the mean is estimated. Therefore, we have that the distribution of α_1 is approximately $\mathcal{N}(-\mu, 3/2m)$. Since the two smaller Gaussians are well separated and equal, the distribution of α_2 is approximately the average of the sample means of the Gaussians, namely $\mathcal{N}(\mu/2, 3/m)$.

Let $\beta = (\alpha_1 + \alpha_2)/2$ be a random variable denoting the border point between the two clusters. Since α_1 and α_2 are approximately independent, we have that the distribution of β is approximately

$\mathcal{N}(-\mu/4, 9/8m)$. As a result, if we let β' and β'' be two independent copies of β, we have that $\beta' - \beta''$ is distributed as $\mathcal{N}(0, 9/4m)$. Finally, since for large values of m we have that β is concentrated around $-\mu/4$, it follows that the probability mass of $\mathcal{D}$ which switches between clusters (over the draw and clustering of two independent samples) is approximately distributed as $|\beta' - \beta''|p(-\mu/4)$, wher $p(\cdot)$ is the probability density function of $\mathcal{D}$. Informally, this is the probability mass which was on 'one side of the border' under the first clustering, and on the 'other side of the border' under the second clustering.

Recall that $d_{\mathcal{D}}(A_k(S_1), A_k(S_2))$ is defined as the probability that two instances sampled from $\mathcal{D}$ will be in the same cluster for clustering $A_k(S_1)$ and in different clusters for clustering $A_k(S_2)$, or vice versa. For $k = 2$ clusters, this reduces to $2t(1-t)$, where t is a random variable defined over a pair of independent samples S_1 and S_2, and represents the probability mass of $\mathcal{D}$ which switches clusters between $A_2(S_1)$ and $A_2(S_2)$. By the results of the previous paragraph, t is distributed as $|\beta' - \beta''|p(-\mu/4)$. Therefore, we have that:

$$\begin{aligned}
stab(A_2, \mathcal{D}, m) &= \mathbb{E}[d_{\mathcal{D}}(A_2(S_1), A_2(S_2))] \\
&= \mathbb{E}[2t(1-t)] \\
&\approx 2\mathbb{E}[p(-\mu/4)|\beta' - \beta''|] - 2\mathbb{E}[(p(-\mu/4))^2(\beta' - \beta'')^2] \\
&\approx \frac{2}{6\sqrt{2\pi}} \exp\left(-\mu^2/32\right) \mathbb{E}[|\beta' - \beta''|] - \frac{2}{72\pi} \exp\left(-\mu^2/16\right) var(\beta' - \beta'') \\
&\approx \frac{1}{3\sqrt{2\pi}} \exp\left(-\mu^2/32\right) \sqrt{\frac{2}{\pi}}\sqrt{\frac{9}{4m}} - \frac{1}{36\pi} \exp\left(-\mu^2/16\right) \frac{9}{4m} \\
&\overset{(1)}{\approx} \frac{1}{2\pi\sqrt{m}} \exp\left(-\mu^2/32\right) \\
&> \frac{1}{7\sqrt{m}} \exp\left(-\mu^2/32\right).
\end{aligned}$$

Step (1) is due to the fact that for large m and/or μ, the second term is negligible compared to the first term.

For $k = 3$ (see figure 1), each centroid is approximately independent and equal to the sample mean of each Gaussian, and therefore the distributions of the two cluster border points β_1 and β_2 are $\mathcal{N}(-\mu/2, 15/8m)$ and $\mathcal{N}(\mu/2, 3/m)$ respectively. Let t_1 denote the probability mass of $\mathcal{D}$ which switches between the two leftmost clusters (over drawing and clustering two independent samples), and let t_2 denote the probability mass of $\mathcal{D}$ which switches between the two rightmost clusters. Since the two leftmost clusters constitute approximately $5/6$ of the sample, and the two rightmost clusters constitute approximately $1/3$ of the sample, we have that the probability that two instances will be in the same cluster under one clustering, and in different clusters under another clustering, is approximately $2t_1(5/6 - t_1) + 2t_2(1/3 - t_2)$. As before, let β_1', β_1'' be two identical independent copies of β_1, and β_2', β_2'' be two identical independent copies of β_2. We have that $\beta_1' - \beta_1''$ is distributed as $\mathcal{N}(0, 15/4m)$ and $\beta_2' - \beta_2''$ is distributed as $\mathcal{N}(0, 6/m)$. Therefore:

$$\begin{aligned}
stab(A_3, \mathcal{D}, m) &= \mathbb{E}[d_{\mathcal{D}}(A_3(S_1), A_3(S_2))] \\
&\approx \mathbb{E}[2t_1(\frac{5}{6} - t_1)] + [\mathbb{E}2t_2(\frac{1}{3} - t_2)] \\
&\approx \frac{5}{3}\mathbb{E}[t_1] + \frac{2}{3}\mathbb{E}[t_2] \\
&\approx \frac{5}{3}p(-\mu/2)\mathbb{E}[|\beta_1' - \beta_1''|] + \frac{2}{3}p(\mu/2)\mathbb{E}[|\beta_2' - \beta_2''|] \\
&\approx \frac{5}{3}\frac{5}{6\sqrt{2\pi}} \exp\left(-\mu^2/8\right) \sqrt{\frac{2}{\pi}}\sqrt{\frac{15}{4m}} + \frac{2}{3}\frac{1}{3\sqrt{2\pi}} \exp\left(-\mu^2/8\right) \sqrt{\frac{2}{\pi}}\sqrt{\frac{6}{m}} \\
&= \sqrt{\frac{6250}{864\pi^2 m}} \exp\left(-\mu^2/8\right) + \sqrt{\frac{8}{27\pi^2 m}} \exp\left(-\mu^2/8\right) \\
&< \frac{1.1}{\sqrt{m}} \exp\left(-\mu^2/8\right).
\end{aligned}$$

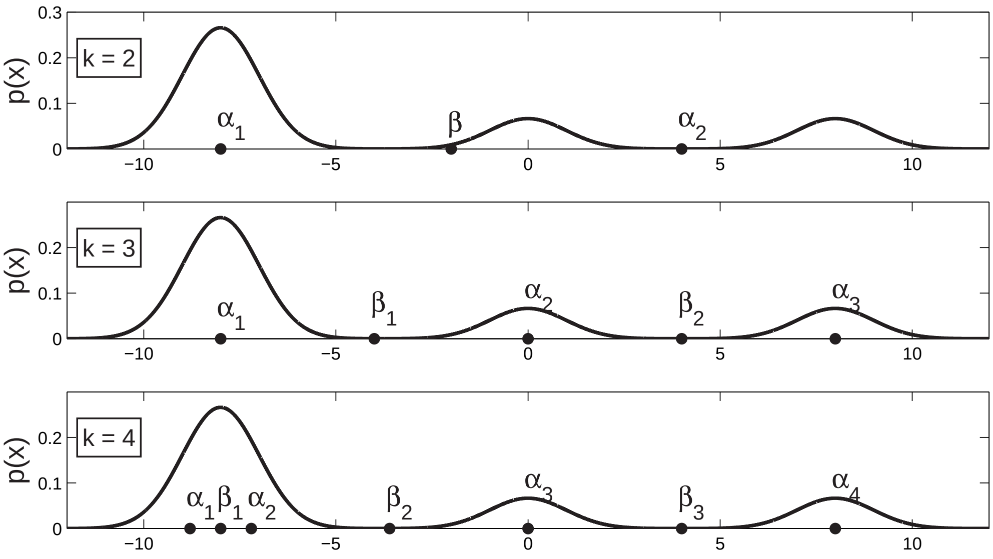

Figure 1: Illustration of centroids and cluster border positions for $k = 2$ (upper sub-figure), $k = 3$ (middle sub-figure), and $k = 4$ (lower sub-figure). The curve represents the density function of $\mathcal{D}$. For large enough sample sizes, the cluster centroids (denoted by α) and cluster border points (denoted by β) will be tightly concentrated around the positions indicated in the sub-figures.

For $k = 4$ (see figure 1), we have two centroids α_1, α_2 on the larger Gaussian, and two centroids α_3, α_4 on the two smaller Gaussians. In this case, the expected probability mass which switches clusters over different samplings is overwhelmingly in the region between the clusters of α_1 and α_2, because all other border areas are in low density areas of $\mathcal{D}$ (taking them into account only improves the derived lower bound).

By theorem 2 in [1], the distribution of β_1 has an asymptotically Gaussian distribution, with a variance which for simplicity will be lower bounded by $3/2m$[1].

As a result, if β_1' and β_1'' are two identical copies of β_1, we have that $\beta_1' - \beta_1''$ is approximately distributed as a Gaussian centered on 0 with a variance of at least $3/m$. We can repeat an argument similar to the other cases (and with the same notation) to get that:

$$\begin{aligned}
stab(A_4, \mathcal{D}, m) &= \mathbb{E}[d_{\mathcal{D}}(A_4(S_1), A_4(S_2))] \\
&\geq \mathbb{E}[2t_1(\frac{2}{3} - t_1)] \\
&\approx \frac{4}{3}\mathbb{E}[t_1] \\
&\approx \frac{4}{3}p(-\mu)\mathbb{E}[|\beta_1' - \beta_1''|] \\
&\geq \frac{8}{3\pi\sqrt{3m}} \\
&> \frac{0.4}{\sqrt{m}}.
\end{aligned}$$

□

[1] In fact, this bound on the variance can be derived directly without resorting to the asymptotic assumption. Since β_1 may be viewed as an unbiased estimator of the larger Gaussian's mean, we can get the result by a direct application of the Crámmer-Rao lower bound.

2 Proof of Lemma 2

Proof. $d_{\mathcal{D}}(A_3(S_1), A_3(S_2))$ is a random variable (over the draw of S_1 and S_2). Its expected value is $stab(A_3, \mathcal{D}, m)$, which by the previous lemma can be upper bounded (up to asymptotically negligible approximation errors) by $1.1\exp(-\mu^2/8)/\sqrt{m}$. Therefore, by Markov's inequality, we have that

$$\Pr\left(d_{\mathcal{D}}(A_3(S_1'), A_3(S_2')) \geq \frac{1}{2\sqrt{m}}\exp(-\mu^2/16)\right) < 2.2\exp(-\mu^2/16). \tag{1}$$

We now wish to prove a lower bound on $d_{\mathcal{D}}(A_2(S_1), A_2(S_2))$ which would hold with high probability. In the proof of lemma 1, we have shown that the distribution of $d_{\mathcal{D}}(A_2(S_1), A_2(S_2))$ is approximately (up to negligible factors) $2p(-\mu/4)|\beta' - \beta''|$, where $\beta' - \beta''$ has a normal distribution $\mathcal{N}(0, 9/4m)$, and $p(\cdot)$ is the probability density function of $\mathcal{D}$. Therefore:

$$\begin{aligned}
&\Pr\left(d_{\mathcal{D}}(A_2(S_1), A_2(S_2)) \;<\; \frac{1}{\sqrt{m}}\exp(-\mu^2/16)\right) \\
&\approx \Pr\left(\frac{1}{3\sqrt{2\pi}}\exp(-\mu^2/32)|\beta' - \beta''| \;<\; \frac{1}{\sqrt{m}}\exp(-\mu^2/16)\right) \\
&= \Pr\left(|\beta' - \beta''| \;<\; \frac{3\sqrt{2\pi}}{\sqrt{m}}\exp(-\mu^2/32)\right) \\
&\overset{(1)}{\approx} 2\Pr\left(\beta' - \beta'' \;<\; \frac{3\sqrt{2\pi}}{\sqrt{m}}\exp(-\mu^2/32)\right) - 1 \\
&\overset{(1)}{\approx} \operatorname{erf}\left(2\sqrt{\pi}\exp(-\mu^2/32)\right) \\
&\overset{(2)}{\leq} 4\exp(-\mu^2/32).
\end{aligned} \tag{2}$$

Step (1) is by the normal distribution of $\beta' - \beta''$ as specified above, and (2) is due to the bound $\operatorname{erf}(x) \leq 2x/\sqrt{\pi}$ for $x \geq 0$.

In the same way, we can derive a high-probability lower bound on $d_{\mathcal{D}}(A_4(S_1), A_4(S_2))$. In the proof of lemma 1, we have shown that the distribution of $d_{\mathcal{D}}(A_4(S_1), A_4(S_2))$ is approximately (up to negligible factors) $(4/3)p(-\mu)|\beta_1' - \beta_1''|$, where $\beta_1' - \beta_1''$ has a normal distribution with variance of at least $3/m$. Repeating the same argument as above, we have that

$$\begin{aligned}
&\Pr\left(d_{\mathcal{D}}(A_4(S_1), A_4(S_2)) \;<\; \frac{1}{\sqrt{m}}\exp(-\mu^2/16)\right) \\
&\approx \Pr\left(\frac{8}{9\sqrt{2\pi}}|\beta_1' - \beta_1''| \;<\; \frac{1}{\sqrt{m}}\exp(-\mu^2/16)\right) \\
&\approx 2\Pr\left(\frac{8}{9\sqrt{2\pi}}(\beta_1' - \beta_1'') \;<\; \frac{1}{\sqrt{m}}\exp(-\mu^2/16)\right) - 1 \\
&= 2\Pr\left(\beta_1' - \beta_1'' \;<\; \frac{9\sqrt{2\pi}}{8\sqrt{m}}\exp(-\mu^2/16)\right) \\
&\leq \operatorname{erf}\left(\frac{3\sqrt{3\pi}}{8}\exp(-\mu^2/16)\right) \\
&\leq \frac{3\sqrt{3}}{4}\exp(-\mu^2/16).
\end{aligned} \tag{3}$$

Combining inequalities 1,2,3, using the union bound, and taking into account the approximations along the way, we have that:

$$\Pr\left(\frac{\min\left\{d_{\mathcal{D}}(A_2(S_1'),A_2(S_2')),d_{\mathcal{D}}(A_4(S_1''),A_4(S_2''))\right\}}{d_{\mathcal{D}}(A_3(S_1),A_3(S_2))}\leq 2\right)$$
$$< (4+o(1))\left(\exp\left(-\frac{\mu^2}{16}\right)+\exp\left(-\frac{\mu^2}{32}\right)\right)$$

□

References

[1] J.A Hartigan. Asymptotic distributions for clustering criteria. *The Annals of Statistics*, 6(1):117–131, 1978.