Critical points of one-dimensional Gaussian mixtures

Aleksander Doan, Dmitry Novikov

7th December 2013

Abstract

We find an optimal upper bound for the number of critical points of a one-dimensional Gaussian mixture. This is the first draft of the paper. In the next version a bound for the small variance case in higher dimensions will be added. The paper is a summary of a summer research project of the first author, supervised by the second author. The project was a part of the Kupcinet-Getz International Summer School of Science 2012 at the Weizmann Institute of Science in Rehovot, Israel.

1 Introduction

Let σ be a positive real number and $(\mu_1, \ldots, \mu_n)$ be n distinct real numbers. We consider the *one-dimensional Gaussian mixture with means* $(\mu_1, \ldots, \mu_n)$ *and variance* σ^2:

$$f_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right).$$

For convenience, we will omit the constant $\sqrt{2\pi}$ and assume that $\mu_1 < \mu_2 < \cdots < \mu_n$. Such functions appear naturally in so-called Gaussian mixture statistical models, which are widely used in applied sciences (see for example [R]).

We will be interested in critical points of such functions. Recall that a point $x_0 \in \mathbb{R}$ is called a *critical point* of f_σ if the derivative f'_σ vanishes at x_0. It is easy to see that all the critical points of f_σ are contained in the compact interval.

Lemma 1. *The function f_σ has no critical points outside the interval $[\mu_1, \mu_n]$.*

Proof. It follows immediately from the equation

$$f'_\sigma(x) = -\frac{1}{\sigma^3} \sum_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right)(x-\mu_i) = 0$$

that a critical point must be a convex combination of $\mu_1, \ldots, \mu_n$. □

In particular, since f_σ is analytic, and therefore has isolated critical points, there are only finitely many of them. We are interested in the following question:

Problem 1. For arbitrary $\sigma > 0$, natural number n, and pairwise distinct real numbers $(\mu_1, \ldots, \mu_n)$, what is the largest possible number of critical points of the Gaussian mixture with variance σ^2 and means $(\mu_1, \ldots, \mu_n)$.

A similar problem for the electrostatic potential in place of the Gaussian mixture was considered in [GNB], and we follow the ideas from this paper to find a bound for a sufficiently small variance σ. Then we give a detailed proof of the following theorem, solving the problem in general case:

Theorem 1. *For any $\sigma > 0$ and $\mu_1, \ldots, \mu_n$, the Gaussian mixture f_σ has no more than $2n-1$ critical points.*

2 Small variance case

First we prove the theorem for sufficiently small σ. When σ tends to zero, the Gaussian mixture f_σ converges to the sum of n Dirac deltas concentrated at points $\mu_1, \ldots, \mu_n$. Therefore, we expect that for sufficiently small $\sigma > 0$, f_σ has exactly one maximum in a small neighbourhood of every point μ_i, and one minimum between every pair of consecutive points μ_i and μ_{i+1}.

However, since a sum of Dirac deltas is not a function, for the sake of proof, we introduce the auxilliary function

$$h_\sigma(x) = \sigma^2 \ln(\sigma f_\sigma(x)).$$

It has the same critical points as f_σ, but, as we prove in the next lemma, its advantage is that for σ tending to zero, it converges to a piecewise smooth function. Namely, consider the function

$$\phi(x) = \min\left\{(x-\mu_1)^2, \ldots, (x-\mu_n)^2\right\}.$$

Note that ϕ fails to be smooth only at $n-1$ points $\gamma_1, \ldots, \gamma_{n-1}$, which, if we assume that $\mu_1 < \cdots < \mu_n$, lying between consecutive points μ_i and μ_{i+1}.

Lemma 2. *For every $x \in \mathbb{R}$, we have*

$$\lim_{\sigma\to 0} h_\sigma(x) = -\frac{1}{2}\phi(x).$$

The convergence is locally uniform in C^2 class outside the set $\{\gamma_1, \ldots, \gamma_{n-1}\}$.

Proof. Choose a compact interval I between γ_i and γ_{i+1}. Then for every $x \in I$, we have $(x-\mu_i)^2 = \min\left\{(x-\mu_1)^2, \ldots, (x-\mu_n)^2\right\}$ and

$$h_\sigma(x) + \frac{1}{2}\phi(x) = \sigma^2 \ln\left(\sum_{j=1}^n \exp\left(-\frac{1}{2\sigma^2}(x-\mu_j)^2\right)\right) + \frac{1}{2}(x-\mu_i)^2 =$$

$$= \sigma^2\left[\ln\left(\sum_{j=1}^n \exp\left(-\frac{1}{2\sigma^2}(x-\mu_j)^2\right)\right) - \ln\left(\exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right)\right)\right] =$$

$$= \sigma \ln\left(\frac{\sum_{j=1}^n \exp(-\frac{1}{2\sigma^2}(x-\mu_j)^2)}{\exp(-\frac{1}{2\sigma^2}(x-\mu_i)^2)}\right) = \sigma \ln\left(\sum_{j=1}^n c_j\right),$$

where each term $c_j = \exp(-\frac{1}{2\sigma^2}[(x-\mu_j)^2 - (x-\mu_i)^2])$ is not greater than 1, and $c_i = 1$. Hence, the sum belongs to the interval $[1, n]$, on which the logarithm is bounded. Therefore, for some constant $C > 0$, we have the following estimate

$$\left|h_\sigma(x) + \frac{1}{2}\phi(x)\right| \le C\sigma^2,$$

which proves the uniform convergence on the interval I. The calculations for the first and second derivative are similar. □

Lemma 3. *Let $I \subset (\mu_i, \mu_{i+1})$ be a compact interval. Then for all sufficiently small σ, the function f_σ is strictly convex on I, and hence has at most one critical point in I.*

Proof. The proof is based on a simple calculation of the second derivative:

$$f'_\sigma(x) = -\frac{1}{\sigma^3}\sum_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right)(x-\mu_i),$$

$$f''_\sigma(x) = -\frac{1}{\sigma^3}\sum_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right)\left[-\frac{1}{\sigma^2}(x-\mu_i)^2+1\right].$$

If x is separated from all points μ_i, then for sufficiently small σ, the expression in the second bracket is strictly negative, hence the second derivative is strictly positive, which proves the lemma. □

Lemma 4. *Let I be a compact interval containing mu_i and none of γ_j. Then for sufficiently small σ, the function f_σ has at most one critical point in I.*

Proof. Since I does not contain any γ_j, by Lemma 2, for sufficiently small σ the second derivative of the function $\sigma \ln(\sigma f_\sigma)$ on I is close to the second derivative of $-\frac{1}{2}\phi$, that is, -1. Hence, the function is concave and has at most one critical point in I. Note, however, that the critical points of f_σ and $\sigma \ln(\sigma f_\sigma)$ coincide. □

Proposition 1. *For sufficiently small σ, the function f_σ has at most $2n-1$ critical points.*

Proof. This is an immediate consequence of the previous lemmas. Choose any partition $a_0 < a_1 < \ldots < a_{2n}$ such that each interval (a_i, a_{i+1}) contains one point from the sequence $\mu_1, \gamma_1, \mu_2, \gamma_2, \ldots, \gamma_{n-1}, \mu_n$. Then, for sufficiently small σ, the function f_σ has at most one critical point in each $[a_i, a_{i+1}]$, and no critical points outside $[a_0, a_{2n}]$. □

3 General case

Having proved Theorem 1 in case of small σ, we strive to extend this result to general case. The crucial observation is that the map $f(x,\sigma) = f_\sigma(x)$, as a function of two variables, satisfies the *heat equation*

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial f}{\partial \sigma}.$$

Denote by $\mathcal{N}(f,\sigma)$ the number of *spatial critical points of f at σ*, that is, critical points of the function $f(\cdot,\sigma) = f_\sigma : \mathbb{R} \to \mathbb{R}$. It is known (rf. [T]) that if f satisfies the heat equation, $\mathcal{N}(f,\sigma)$ is a non-increasing function of σ. This finishes the proof of the theorem in general case, since for arbitrary σ, we can choose $\sigma_0 < \sigma$ small enough that

$$\mathcal{N}(f,\sigma) \leq \mathcal{N}(f,\sigma_0) \leq 2n-1.$$

We present here an alternative proof of Theorem 1, which does not make use of the non-trivial result from [T]. Here is an outline of the proof:

1. Fix $\sigma > 0$. We want to show that $\mathcal{N}(f,\sigma) \leq 2n-1$.
2. Prove that for a *generic* solution of the heat equation the number of spatial critical points in non-increasing in time (which is much easier than general case).

3. Perturb f to obtain a generic solution $\tilde{f}$. The perturbation has to satisfy two conditions:
   (a) It does not decrease the number of spatial critical at σ.
   (b) The asymptotic behaviour of $\tilde{f}_\sigma$ for $\sigma \to 0$ is the same as for f_σ.
4. Previous points guarantee that for σ_0 small enough we have
$$\mathcal{N}(f,\sigma) \le \mathcal{N}(\tilde{f},\sigma) \le \mathcal{N}(\tilde{f},\sigma_0) \le 2n-1.$$

In Proposition 2 below we investigate the behaviour of a generic solution, whereas in Propositions 3 and 4 we prove that such a perturbation is possible. We conclude then with the formal proof of the Theorem 1.

First we have to set some convenient notation and terminology. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be an analytic function of two variables (x,t). Recall that (x_0,t_0) is a *critical point* of f if $\nabla f(x_0,t_0) = 0$, and x_0 is a *spatial critical point of f at t_0* if the derivative of the one-variable function $f(\cdot,t_0)$ vanishes at x_0.

We are interested in the behaviour of the number $\mathcal{N}(f,t)$ of spatial critical points when the time variable t changes. Consider the real algebraic *curve of spatial critical points $C \subset \mathbb{R}^2$* given by the equation $f_x = 0$. Geometrically, $\mathcal{N}(f,t_0)$ is the total number of all intersection points of C with the line $t = t_0$ in $\mathbb{R}^2$. We will investigate how the number of intersection points changes locally. Let x_0 be a spatial critical point of f at t_0. Choose small $\delta > 0$ and a small neighbourhood U of x_0 in $\mathbb{R}$. Denote by $n(t)$ the number of those spatial critical points of f at t which are contained in U. We introduce the following notions[1]:

1. If $n(t) = n(t_0)$ for all $t \in [t_0-\delta, t_0+\delta]$, then (x_0,t_0) is a *neutral point.*
2. If $n(t-\epsilon) < n(t_0) < n(t_0+\epsilon)$ for all $\epsilon < \delta$, then (x_0,t_0) is a *creation point.*
3. If $n(t-\epsilon) > n(t_0) > n(t_0+\epsilon)$ for all $\epsilon < \delta$, then (x_0,t_0) is an *anihilation point.*

If $(x_0,t_0) \in C$ and $f_{xx}(x_0,t_0) \neq 0$, then by the implicit function theorem, we can find a parametrisation $x = x(t)$ around (x_0,t_0) such that $f_x(x,t) = 0$ if and only if $x = x(t)$, which means that (x_0,t_0) is a neutral point. Therefore, we are interested in spatial critical points $(x_0,t_0) \in C$ such that $f_{xx}(x_0,t_0) = 0$. Let us call them *essential critical points.*

Proposition 2. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a solution of the heat equation $f_{xx} = f_t$. Then a point x_0 is an essential critical point of $f(\cdot,t_0)$ if and only if (x_0,t_0) is a critical point of f as a function of two variables. Moreover, non-degeneracy of (x_0,t_0) as a critical point of f is equivalent to the condition $f_{xt}(x_0,t_0) \neq 0$ and in this case (x_0,t_0) is an anihilation point.*

Proof. Compute the gradient and Hessian matrix of f:
$$\nabla f = \begin{pmatrix} f_x \\ f_t \end{pmatrix} = \begin{pmatrix} f_x \\ f_{xx} \end{pmatrix},$$
$$H(f) = \begin{pmatrix} f_{xx} & f_{xt} \\ f_{xt} & f_{tt} \end{pmatrix}.$$
Therefore, $\nabla f = 0$ if and only if $f_x = f_{xx} = 0$ and in this case $\det H(f) = -f_{xt}^2$, which proves the first part of the proposition.

[1]Of course, these are not the only possibilities.

Assume now that $f_{xt}(x_0, t_0) \neq 0$. Then, by the implicit function theorem, we can find a parametrisation $t = t(x)$ of the set $\{f_x = 0\}$ around (x_0, t_0). Differentiating twice the equation $f_x(x, t(x)) = 0$ with respect to x and denoting $t' = dt/dx$, we obtain

$$f_{xx} + f_{xt}t' = 0,$$
$$f_{xxx} + f_{xxt}t' + f_{xxt}t' + f_{xtt}(t')^2 + f_{xt}t'' = 0.$$

Setting now $x = x_0$ and using equalities $t(x_0) = t_0$, $f_{xx}(x_0, t_0) = 0$, we get from the first equation that $f_{xt}(x_0, t_0)t'(x_0) = 0$, which means $t'(x_0) = 0$, because $f_{xt}(x_0, t_0)$. Then the second equation yields

$$f_{xt}(x_0, t_0)\,(1 + f''(x_0)) = f_{xxx}(x_0, t_0) + f_{xt}(x_0, t_0)t''(x_0) = 0,$$

where we have used the heat equation $f_{xxx} = f_{xt}$. Therefore, we obtain $t'(x_0) = 0$ and $t''(x_0) = -1$, which means that the graph of the function $t(x)$ is tangent to the horizontal line $t = t_0$ at x_0 and concave in its neighbourhood. Hence, around x_0 it must stay below the line $t = t_0$, so the critical points of $f(\cdot, t)$ for $t < t_0$ are anihilated and for $t > t_0$ there are no critical points around x_0.

□

Proposition 3. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a solution to the heat equation $f_{xx} = f_t$. Then for almost all $c = (c_1, c_2, c_3) \in \mathbb{R}^3$, the function*

$$f^c(x, t) = f(x, t) + c_1 x + c_2(x^2 + 2t) + c_3(x^3 + 6tx)$$

is a solution with only non-degenerate critical points.

Proof. It is easy to check by direct calculation that f^c satisfies the heat equation. To prove that for a generic c, all the critical points are non-degenerate, we shall use the Thom transversality theorem.

Consider the mapping

$$F : \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}^3,$$
$$F(x, t, c) = (f^c_x, f^c_{xx}, f^c_{xt})\,.$$

For fixed $c \in \mathbb{R}^3$, the set $Z_c = \{(x, t) \mid F(x, t, c) = 0\}$ consists precisely of degenerate critical points of f^c (see the proof of Proposition 2). If we manage to show that for almost all $c \in \mathbb{R}^3$, the map $F(\cdot, c)$ is transversal to the zero-dimensional submanifold $0 \in \mathbb{R}^3$, then it follows from dimensional reasons that for almost all c, the set Z_c is empty, and therefore f^c has no non-degenerate critical points at all.

By transversality theorem, it suffices to show that F is transversal to $0 \in \mathbb{R}^3$ as a map from $\mathbb{R}^2 \times \mathbb{R}^3$ to $\mathbb{R}^3$. By simple calculation, we easily see that its derivative $\nabla F(x, t, c)$, which is represented by a 5×3 matrix, contains the square matrix

$$\begin{pmatrix} f^c_{xc_1} & f^c_{xc_2} & f^c_{xc_3} \\ f^c_{xxc_1} & f^c_{xxc_2} & f^c_{xxc_3} \\ f^c_{xtc_1} & f^c_{xtc_2} & f^c_{xtc_3} \end{pmatrix} = \begin{pmatrix} 1 & * & * \\ 0 & 2 & * \\ 0 & 0 & 6 \end{pmatrix}$$

of rank 3. Therefore, $F : \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}^3$ is a submersion, which guarantees that the assumptions of the transversality theorem are satisfied. □

Lemma 5. *Let g and $u : \mathbb{R} \to \mathbb{R}$ be analytic functions. Assume that g has only finitely many zeroes. Then there is an interval I of the form $(0, \epsilon)$ or $(-\epsilon, 0)$ for some $\epsilon > 0$, such that for all $\lambda \in I$ the function*

$$g_\lambda = g + \lambda u$$

has at least as many zeroes as g.

Proof. We will investigate the behaviour of the perturbed function g_λ around zeroes of the original function g. If x_0 is a zero of both g and u, then g_λ also vanishes at x_0 for all λ, so it does not affect the difference between the number of zeroes of g and g_λ. Therefore, since now we will assume that u and g have no common zeroes.

Let x_0 be a zero of g of order m, that is,

$$g(x) = (x - x_0)^m h(x)$$

for an analytic function h such that $h(x_0) \neq 0$. Assume that we want to perturb g by a function u which does not vanish at x_0. Without loss of generality, assume that $u(x_0) > 0$. Consider

$$g_\lambda(x) = g(x) + \lambda u(x) = (x - x_0)^m h(x) + \lambda u(x).$$

If $m = 2k+1$ is odd, then for every sufficiently small $|\lambda|$ and $\epsilon > 0$, the values of g_λ at $x_0 - \epsilon$ and $x_0 + \epsilon$ are of different signs, hence g_λ has a zero in the interval $(x_0 - \epsilon, x_0 + \epsilon)$.

If $m = 2k$ is even, then consider the equation

$$(x - x_0)^{2k} h(x) + \lambda u(x) = 0. \tag{1}$$

Note that in this case x_0 is either a local minimum, or local maximum, depending on the sign of $h(x_0)$. Since the left hand side has the partial derivative with respect to λ not vanishing at the point $(x, \lambda) = (x_0, 0)$, by the implicit function theorem, on a neighbourhood of x_0 there exists an analytic function $\lambda(x)$ with $\lambda(x_0) = 0$ such that the above equation is satisfied if and only if $\lambda = \lambda(x)$. Differentiating the equation $2n$ times, we immediately obtain

$$\lambda'(x_0) = \lambda''(x_0) = \cdots = \lambda^{(2n-1)}(x_0) = 0,$$

$$(2k)!h(x_0) + \lambda^{(2n)}(x_0)u(x_0) = 0.$$

Hence, $\lambda^{2n}(x_0) \neq 0$ and λ is of the form

$$\lambda(x) = (x - x_0)^{2n} w(x),$$

for some function w with $w(x_0) \neq 0$. The sign of $w(x_0)$ depends only on the sign of $h(x_0)$, since we have assumed that $u(x_0) > 0$. Now, we easily see that in a small neighbourhood of x_0, for every sufficiently small constant a, the equation

$$(x - x_0)^{2n} w(x) = a$$

has no roots if a has different sign than $w(x_0)$, and at least two if the signs are the same. Indeed, let say that $w(x_0) > 0$ and choose a neighbourhood of x_0 small enough for w to be positive on it. Then for a small $\delta > 0$, the values of λ at $x_0 - \delta$ and $x_0 + \delta$ are positive, so by continuity and $\lambda(x_0) = 0$, λ attains all values from the interval $(0, \min\{\lambda(x_0 - \delta), \lambda(x_0 + \delta)\}]$ at least twice, once in $(x_0 - \delta, x_0)$ and once in $(x_0, x_0 + \delta)$.

Therefore, there is an open interval with 0 as its limit point such that for all a from this interval, the equation $\lambda(x) = a$ has at least two solutions in a neighbourhood of x_0. This means, by definition of $\lambda(x)$, that for all λ belonging to this interval, the equation 1 has at least two solutions, corresponding to zeroes of g_λ. Conversely, for small λ of the opposite sings, the equation has no solutions.

Now, assume that g has l zeroes of odd multiplicity and $k = k_+ + k_-$ zeroes of even multiplicity, where k_+ and k_- denote, respectively, the number of local maxima and local minima amongst these zeroes. Without loss of generality, let us assume that $k_+ \geq k_-$. By the above reasoning,

$g + \lambda u$ for sufficiently small $\lambda \neq 0$ will have at least one zeroe around each of l zeroes of odd multiplicity. Depending on the sign of λ and their type (maximum or minimum), zeroes of even multiplicity will be either anihilated or at least doubled, so that when we choose the sign λ carefully, the total number of zeroes of $g + \lambda u$ will be at least

$$l + 2k_+ \geq l + k_- + k_+,$$

so it will not decrease. □

Proposition 4. *Then there is an open subset $U \subset \mathbb{R}^3$ having $0 \in \mathbb{R}^3$ as a limit point, such that for all $c = (c_1, c_2, c_3) \in U$, the function*

$$g^c(x) = g(x) + c_1 + c_2 x + c_3 x^2$$

has at least as many zeroes as g does.

Proof. The statement is an immediate consequence of Lemma 5. □

Proof of the Theorem 1. We want to show that for an arbitrary $\sigma_0 > 0$, the function f_σ has at most $2n - 1$ critical points.

Step 1. As in Proposition 1, choose a small $\sigma < \sigma_0$ and $a_1 < \ldots < a_{2n}$ such that f_σ is strictly convex in the intervals (a_{2i-1}, a_{2i}) containing the points μ_i, and $\sigma \ln(\sigma f_\sigma)$ is strictly concave in intervals (a_{2i}, a_{2i+1}) containing the points γ_i.

Step 2. Choose $c \in \mathbb{R}^3$ such that f_c has only non-degenerate critical points (Proposition **??**), $(f_c)_{\sigma_0}$ has no fewer critical points than f_{σ_0} (Proposition **??**), and, finally, c is so small that the functions $(f_c)_\sigma$ and $\sigma \ln(\sigma (f_c)_\sigma)$ approximate f_σ and $\sigma \ln(\sigma f_\sigma)$, respectively, in C^2 norm on $[a_1, a_{2n}]$. In particular, they are, respectively, strictly convex and strictly concave on the same intervals as the non-perturbed functions.

Step 3. For a function $g : \mathbb{R}^2 \to \mathbb{R}$ and $t \in \mathbb{R}$ denote by $\mathcal{N}(g, t)$ the number of spatial critical points of $g(\cdot, t)$. Then by the choice of σ and c, we obtain

$$\mathcal{N}(f, \sigma_0) \leq \mathcal{N}(f_c, \sigma_0) \leq \mathcal{N}(f_c, \sigma),$$

since f_c has only non-degenerate critical points and hence, by Proposition 2, the number of critical points is non-increasing with σ. Finally, f_c has at σ at most $2n-1$ spatial critical points, by exactly the same convexity argument which was used in the proof of Proposition 1. Therefore,

$$\mathcal{N}(f, \sigma_0) \leq 2n - 1.$$

□

References

[GNB] A. Gabrielov, D. Novikov, B. Shapiro, *Mystery of point charges*, Proc. Lond. Math. Soc., vol. 95 (2007), pp. 443–472.

[T] L. Turyn, *Spatial critical points of solutions of a one-dimensional nonlinear parabolic problem*, Proc. Amer. Math. Soc., vol. 106 (1989), pp. 1003–1009.

[R] D. A. Reynolds, *Gaussian Mixture Models* in: *Encyclopedia of Biometrics*, Springer US, 2009.