Neural Processing Letters manuscript No.
(will be inserted by the editor)

Gradient-based training of Gaussian Mixture Models for High-Dimensional Streaming Data

Alexander Gepperth · **Benedikt Pfülb**

Received: date / Accepted: date

Abstract We present an approach for efficiently training Gaussian Mixture Model (GMM) by Stochastic Gradient Descent (SGD) with non-stationary, high-dimensional streaming data. Our training scheme does not require data-driven parameter initialization (e.g., k-means) and can thus be trained based on a random initialization. Furthermore, the approach allows mini-batch sizes as low as 1, which are typical for streaming-data settings. Major problems in such settings are undesirable local optima during early training phases and numerical instabilities due to high data dimensionalities. We introduce an adaptive annealing procedure to address the first problem, whereas numerical instabilities are eliminated by using an exponential-free approximation to the standard GMM log-likelihood. Experiments on a variety of visual and non-visual benchmarks show that our SGD approach can be trained completely without, for instance, k-means based centroid initialization. It also compares favorably to an online variant of Expectation-Maximization (EM) – stochastic EM (sEM), which it outperforms by a large margin for very high-dimensional data.

Keywords Gaussian Mixture Model · Stochastic Gradient Descent

1 Introduction

This contribution focuses on Gaussian Mixture Model (GMM), which represent a probabilistic unsupervised model for clustering and density estimation, allowing sampling and outlier detection. GMMs have been used in a wide range of scenarios, see [**?**]. Commonly, free parameters of a GMM are estimated using the Expectation-Maximization (EM) algorithm [**?**], which does not require learning rates and automatically enforces all GMM constraints. A popular online variant is stochachstic EM [**?**], which can be trained mini-batch wise and is thus more suited for large datasets or streaming data.

Fulda University of Applied Sciences
Leipziger Str. 123, 36037 Fulda
E-mail: {alexander.gepperth,benedikt.pfuelb}@cs.hs-fulda.de

1.1 Motivation

Intrinsically, EM is a batch-type algorithm. Therefore, memory requirements can become excessive for large datasets. In addition, streaming-data scenarios require data samples to be processed one by one, which is impossible for a batch-type algorithm. Moreover, data statistics may be subject to changes over time (concept drift/shift), to which the GMM should adapt. In such scenarios, an online, mini-batch type of optimization such as SGD is attractive since it can process samples one by one, has modest, fixed memory requirements, and can adapt to changing data statistics.

1.2 Related Work

Online EM is a technique for performing EM mini-batch wise, allowing to process large datasets. One branch of previous research [**?**,**?**,**?**] has been devoted to the development of stochastic Expectation-Maximization (sEM) algorithms that reduce to the original EM method in the limit of large batch sizes. The variant presented in [**?**] is widely used due to its simplicity and efficiency for large datasets. Such approaches come at the price of additional hyper-parameters (e.g., step size, mini-batch size, step size reduction), thus, removing a key advantage of EM over SGD. Another approach is to modify the EM algorithm itself by, e.g., including heuristics for adding, splitting and merging centroids [**?**,**?**,**?**,**?**,**?**,**?**,**?**]. This allows GMM-like models to be trained by presenting one sample after another. The models work well in several application scenarios, but their learning dynamics are impossible to analyze mathematically. They also introduce a high number of parameters. Apart from these works, some authors avoid the issue of extensive datasets by determining smaller "core sets" of representative samples and performing vanilla EM [**?**].
SGD for training GMM has, as far as we know, been recently treated only in [**?**,**?**]. In this body of work, GMM constraint enforcement is ensured by using manifold optimization techniques and re-parameterization/regularization, thereby introducing additional hyper-parameters. The issue of local optima is side-stepped by a k-means type centroid initialization, and the used datasets are low-dimensional (36 dimensions).
Annealing and Approximation approaches for GMMs were proposed in [**?**,**?**,**?**,**?**]. However, the regularizers proposed in [**?**,**?**] significantly differ from our scheme. GMM log-likelihood approximations, similar to the one used here, are discussed in, e.g., [**?**] and [**?**], but only in combination with EM training. A similar "hard assignment" approximation is performed in [**?**].
GMM Training in High-Dimensional Spaces is discussed in several publications: A conceptually very interesting procedure is proposed in [**?**]. It exploits the properties of high-dimensional spaces in order to achieve learning with a number of samples that is polynomial in the number of Gaussian components. This is difficult to apply in streaming settings, since higher-order moments need to be estimated beforehand, and also because the number of samples is usually unknown. Training GMM-like lower-dimensional factor analysis models by SGD on high-dimensional image data is successfully demonstrated in [**?**]. They avoid numerical issues, but, again, sidestep the local optima issue by using k-means initialization. The numerical issues associated with log-likelihood computation in high-dimensional spaces

are generally mitigated by using the "logsumexp" trick [**?**], which is, however, insufficient for ensuring numerical stability for particularly high-dimensional data, such as images.

1.3 Goals and Contributions

The goals of this article are to establish GMM training by SGD as a simple and scalable alternative to sEM in streaming scenarios with potentially high-dimensional data. The main novel contributions are:

- a proposal for numerically stable GMM training by SGD that outperforms sEM for high data dimensionalities,
- an automatic annealing procedure that ensures SGD convergence without prior knowledge of the data (**no** k-means initialization) which is beneficial for streaming data,
- a computationally efficient method for enforcing all GMM constraints in SGD,
- a convergence proof for the annealing procedure.

Additionally, we provide a TensorFlow implementation.[1]

2 Gaussian Mixture Models

GMMs are probabilistic models that intend to explain the observed data $X = \{\boldsymbol{x}_n\}$ by expressing their density as a weighted mixture of K Gaussian component densities $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{P}_k) \equiv \mathcal{N}_k(\boldsymbol{x})$:

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}_k(\boldsymbol{x}_n). \tag{1}$$

Here, we parameterize Gaussian densities by precision matrices $\boldsymbol{P}_k = \boldsymbol{\Sigma}_k^{-1}$ instead of covariances $\boldsymbol{\Sigma}_k$. The component weights π_k represent another set of GMM parameters, which modulate the overall influence of the Gaussian distribution. For a derivation of eq. (1), we must introduce the probabilistic foundations of GMMs. These models assume that each observed data sample $\{\boldsymbol{x}_n\}$ is drawn from one of the Gaussian component densities $\mathcal{N}_k$. The selection of this component density is assumed to depend on an unobserved (and unobservable) *latent variable* $z_n \in \{1, \ldots, K\}$ which follows an unknown distribution. This is formalized for a GMM with K components by formulating the *complete-data likelihood* for a single data sample as:

$$p(\boldsymbol{x}_n, z_n) = \pi_{z_n} \mathcal{N}_{z_n}(\boldsymbol{x}_n), \tag{2}$$

Since the latent variables are, by construction, unobservable, we must marginalize them out of eq. (15) in order to obtain an expression suitable for optimization. This gives us the *incomplete-data likelihood* for a single data sample $\boldsymbol{x}_n$:

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} p(\boldsymbol{x}_n, k), \tag{3}$$

[1] https://gitlab.cs.hs-fulda.de/ML-Projects/sgd-gmm

which depends on observable quantities only. Please compare this to eq. (1). The incomplete-data likelihood for all samples is thus given by:

$$p(X) = \prod_n p(\boldsymbol{x}_n) = \prod_n \sum_k p(\boldsymbol{x}_n, k) = \prod_n \sum_k \pi_k \mathcal{N}_k(\boldsymbol{x}_n), \tag{4}$$

where we have inserted eq. (15) in the last step. Passing to the log-domain (as it is common for probabilistic models), we obtain the *total incomplete-data log-likelihood* for all observed data samples:

$$\mathcal{L} = \log p(X) = \sum_n \log \sum_k \pi_k \mathcal{N}_k(\boldsymbol{x}_n). \tag{5}$$

The function $\mathcal{L}$ contains only observable quantities and is a suitable loss function for optimization. For convenience and numerical stability, the sum is usually replaced by an expectation, and we follow this convention:

$$\mathcal{L} = \mathbb{E}_n \left[\log \sum_k \pi_k \mathcal{N}_k(\boldsymbol{x}_n) \right]. \tag{6}$$

Please note that $\mathcal{L}$ represents the likelihood of the observed data under the GMM with current parameters, and must therefore be *maximized* to obtain a better explanation of the data.

2.1 GMM Constraint Enforcement for SGD

GMMs require the mixture weights to be normalized: $\sum_k \pi_k = 1$ and the precision matrices to be positive definite: $\boldsymbol{x}^\top \boldsymbol{P}_k \boldsymbol{x} \geq 0 \; \forall \boldsymbol{x}$. These constraints must be explicitly enforced after each SGD step:
Weights π_k are adapted according to [**?**], which replaces them by other free parameters ξ_k from which the π_k are computed so that normalization is ensured:

$$\pi_k = \frac{\exp(\xi_k)}{\sum_j \exp(\xi_j)}. \tag{7}$$

Diagonal precision matrices are re-parameterized as $\boldsymbol{P}_k = \boldsymbol{D}_k^2$, with diagonal matrices $\boldsymbol{D}_k$ (Cholesky decomposition). They are, therefore, guaranteed to be positive definite. Hence, $\det \boldsymbol{\Sigma}_k = \det \boldsymbol{P}_k^{-1} = \big(\det(\boldsymbol{D}_k^2)\big)^{-1} = \big(\operatorname{Tr}(\boldsymbol{D}_k)\big)^{-2}$ can be computed efficiently. Since we are dealing with high-dimensional data, precision matrices are always taken to be diagonal, since full matrices would be prohibitive w.r.t. memory consumption and the number of free parameters.
Full precision matrices are treated here for completeness' sake, since they are infeasible for high-dimensional data. We represent them as a spectral decomposition into eigenvectors $\boldsymbol{v}_i$ and eigenvalues λ_i^2: $\boldsymbol{P}_k = \sum_i \lambda_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^\top$, which ensures positive-definiteness. This can be seen from $\det \boldsymbol{\Sigma}_k = \det \boldsymbol{P}_k^{-1} = \prod_i \lambda_i^{-2}$. In order to maintain a correct representation of eigenvectors, these have to be orthonormalized after each SGD step.

2.2 Max-Component Approximation for GMM

The log-likelihood eq. (5) is difficult to optimize by SGD due to numerical problems (mainly underflows and resulting divisions by zero) for high data dimensionalities. This is why we intend to find a lower bound that we can optimize instead. A simple scheme is given by

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_n\left[\log \sum_k \pi_k \mathcal{N}_k(\boldsymbol{x}_n)\right] \geq \mathbb{E}_n\left[\log \max_k\left(\pi_k \mathcal{N}_k(\boldsymbol{x}_n)\right)\right] \\ &= \hat{\mathcal{L}} = \mathbb{E}_n\left[\log\left(\pi_{k^*} \mathcal{N}_{k^*}(\boldsymbol{x}_n)\right)\right]\end{aligned} \tag{8}$$

where $k^* = \arg\max_k \pi_k \mathcal{N}_k(\boldsymbol{x}_n)$. This is what we call the *max-component approximation* of eq. (8). In contrast to the lower bound that is constructed for EM-type algorithms, our bound is usually not tight. Nevertheless, we will demonstrate later that it is a very good approximation when data are high-dimensional. The advantage of $\hat{\mathcal{L}}$ is the elimination of exponentials causing numerical instabilities. The "logsumexp" trick is normally employed with GMMs to rectify this by factoring out the largest component probability $\mathcal{N}_{k^*}$. This mitigates but does not avoid numerical problems when distances are high, a common occurrence for high data dimensions. To give an example: we normalize the component probability $\mathcal{N}_k = e^{-101}$ (using 32-bit floats) by the highest probability $\mathcal{N}_{k^*} = e^3$, and we obtain $\frac{\mathcal{N}_k}{\mathcal{N}_{k^*}} = e^{-104}$, which produces an underflow.

2.3 Undesirable Local Optima in SGD Training

A crucial issue when optimizing $\hat{\mathcal{L}}$ (and indeed $\mathcal{L}$ as well) by SGD without k-means initialization concerns undesirable local optima. Most notable are the **single/sparse-component solutions**, see fig. 1. They are characterized by one or several components $\{k_i\}$ having large weights, with centroid and precision matrices given by the mean and covariance of a significant subset $\boldsymbol{X}_{k_i} \subset \boldsymbol{X}$ of the data $\boldsymbol{X}$: $\pi_{k_i} \gg 0$, $\boldsymbol{\mu}_{k_i} = \mathbb{E}[\boldsymbol{X}_{k_i}]$, $\boldsymbol{\Sigma}_{k_i} = \mathrm{Cov}(\boldsymbol{X}_{k_i})$, whereas the remaining components k are characterized by $\pi_k \approx 0$, $\boldsymbol{\mu}_k = \boldsymbol{\mu}(t=0)$, $\boldsymbol{P}_k = \boldsymbol{P}(t=0)$. Thus, these unconverged components are almost never Best Matching Unit (BMU) k^*. The max-operation in $\hat{\mathcal{L}}$ causes gradients like $\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{\mu}_k}$ to contain δ_{kk^*}:

$$\begin{aligned}\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{\mu}_k} &= \mathbb{E}_n\left[\boldsymbol{P}_k\left(\boldsymbol{x}_n - \boldsymbol{\mu}_k\right)\delta_{kk^*}\right] \\ \frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{P}_k} &= \mathbb{E}_n\left[\left(\left(\boldsymbol{P}_k\right)^{-1} - \left(\boldsymbol{x}_n - \boldsymbol{\mu}_k\right)\left(\boldsymbol{x}_n - \boldsymbol{\mu}_k\right)^\top\right)\delta_{kk^*}\right] \\ \frac{\partial \hat{\mathcal{L}}}{\partial \pi_k} &= \pi_k^{-1}\mathbb{E}_n\left[\delta_{kk^*}\right].\end{aligned} \tag{9}$$

This implies that the gradients are non-zero only for the BMU k^*. Thus, the gradients of unconverged components vanish, implying that they remain in their unconverged state.

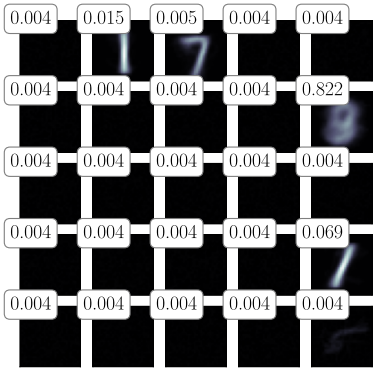


Fig. 1 A sparse-component-solution with superimposed component weights π_k, obtained when performing naive SGD on MNIST.

2.4 Annealing Procedure for Avoiding Local Optima

Our approach for avoiding sparse-component solutions is to punish their characteristic response patterns by replacing $\hat{\mathcal{L}}$ by the *smoothed max-component log-likelihood* $\hat{\mathcal{L}}^\sigma$:

$$\begin{aligned}\hat{\mathcal{L}}^\sigma &= \mathbb{E}_n \max_k \left[\sum_j \boldsymbol{g}_{kj}(\sigma) \log\Big(\pi_j \mathcal{N}_j(\boldsymbol{x}_n)\Big) \right] \\ &= \mathbb{E}_n \sum_j \boldsymbol{g}_{k^* j}(\sigma) \log\Big(\pi_j \mathcal{N}_j(\boldsymbol{x}_n)\Big).\end{aligned} \tag{10}$$

Regarding its interpretation, we are assuming that the K GMM components are arranged in a quadratic 2D grid of size $\sqrt{K} \times \sqrt{K}$. Equally, each $\boldsymbol{g}_k$ is interpreted as 2D grid of size $\sqrt{K} \times \sqrt{K}$, (see fig. 2), with values given by a periodically continued 2D Gaussian centered on component k. With this interpretation, Equation (10) represents a 2D convolution with periodic boundary conditions (in the sense used in image processing) of the $\log(\pi_k \mathcal{N}_k(\boldsymbol{x}))$ by a smoothing filter whose width is controlled by σ. Thus, eq. (10) is maximized if the log-probabilities follow a uni-modal Gaussian profile of spatial variance $\sim \sigma^2$, which heavily punishes single/sparse-component solutions that have a locally delta-like response. A 1D grid for annealing, together with 1D smoothing filters, was verified to fulfill this purpose as well. We chose 2D because it allows for an easier visualization while incurring an identical computational cost.

Annealing starts with a large value of $\sigma(t) = \sigma_0$ and reduces it over time to an asymptotic small value of $\sigma = \sigma_\infty$, thus, smoothly transitioning from $\hat{\mathcal{L}}^\sigma$ in eq. (10) into $\hat{\mathcal{L}}$ in eq. (8).

Annealing Control regulates the decrease of σ. This quantity defines an effective upper bound on $\hat{\mathcal{L}}^\sigma$ (see section 2.6 for a proof). An implication is that the loss will be stationary once this bound is reached, which we consider a suitable indicator for reducing σ. We implement an annealing control that sets $\sigma \leftarrow 0.9\sigma$ whenever the loss is considered sufficiently stationary. Stationarity is detected by maintaining an exponentially smoothed average $\ell(t) = (1-\alpha)\ell(t-1) + \alpha\hat{\mathcal{L}}^\sigma(t)$ on

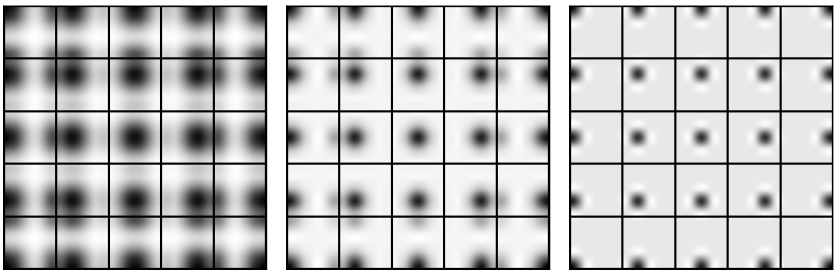


Fig. 2 Visualization of Gaussian smoothing filters $\boldsymbol{g}_k$, of width σ, used in annealing for three different values of σ. The $\boldsymbol{g}_k$ are placed on a 2D grid, darker pixels indicate larger values. Over time, $\sigma(t)$ is reduced (middle and right pictures) and the Gaussians approach a delta peak, thus, recovering the original, non-annealed loss function. Note that the grid is considered periodic in order to avoid boundary effects, so the $\boldsymbol{g}_k$ are themselves periodic.

time scale α. Every $\frac{1}{\alpha}$ iterations, we compute the fractional increase of $\hat{\mathcal{L}}^\sigma$ as

$$\Delta = \frac{\ell(t) - \ell(t - \alpha^{-1})}{\ell(t - \alpha^{-1}) - \hat{\mathcal{L}}^\sigma(t = 0)} \tag{11}$$

and consider the loss stationary iff $\Delta < \delta$ (the latter being a free parameter). The choice of the time constant for smoothing $\hat{\mathcal{L}}^\sigma$ is outlined in the following section.

2.5 Training Procedure for SGD

Training GMMs by SGD is performed by maximizing the smoothed max-component log-likelihood $\hat{\mathcal{L}}^\sigma$ from eq. (10). At the same time, we enforce the constraints on the component weights and covariances as described in section 2.1 and transition from $\hat{\mathcal{L}}^\sigma$ into $\hat{\mathcal{L}}$ by annealing (see section 2.4). SGD requires a learning rate ϵ to be set, which in turn determines the parameter α (see section 2.4) as $\alpha = \epsilon$ since stationarity detection should operate on a time scale similar to that of SGD. The diagonal matrices $\boldsymbol{D}_k$ are initialized to $D_{\max} I$ and are clipped after each iteration so that diagonal entries remain in the range $[0, D_{\max}^2]$. This is necessary to avoid excessive growth of precisions for data entries with vanishing variance, e.g., pixels that are always black. Weights are uniformly initialized to $\pi^i = \frac{1}{K}$, centroids in the range $[-\mu^i, +\mu^i]$ (see algorithm 1 for a summary). Please note that our SGD approach requires no centroid initialization by k-means, as it is recommended when training GMMs with (s)EM. We discuss and summarize good practices for choosing hyper-parameters in section 5.

2.6 Proof that Annealing is Convergent

We assume that, for a fixed value of σ, SGD optimization has reached a stationary point where the derivative w.r.t. all GMM parameters is 0 on average. In this situation, we claim that decreasing σ will always increase the loss. If true, this would show that σ defines an effective upper bound for the loss. For this to be consistent, we have to show that the loss gradient w.r.t. σ vanishes as $\sigma \to 0$: as the annealed loss approaches the original one, decreases of σ have less and less effects.

Algorithm 1: Steps of SGD-GMM training.

Data: initializer values: μ^i, K, $\epsilon_0/\epsilon_\infty$, σ_0/σ_∞, δ and data $\boldsymbol{X}$
Result: trained GMM model
$\boldsymbol{\mu} \leftarrow \mathcal{U}(-\mu^i, +\mu^i)$, $\pi \leftarrow 1/K$, $\boldsymbol{P} \leftarrow ID_{\max}$, $\sigma \leftarrow \sigma_0$, $\epsilon \leftarrow \epsilon_0$
forall $t < T$ **do** // training loop
  $\boldsymbol{g}(t) \leftarrow$ create_annealing_mask(σ,t) // see section 2.4
  $\boldsymbol{\mu}(t) \leftarrow \epsilon \frac{\partial \hat{\mathcal{L}}^\sigma}{\partial \boldsymbol{\mu}} + \boldsymbol{\mu}(t\text{-}1)$, // SGD updates
  $\boldsymbol{P}(t) \leftarrow \epsilon \frac{\partial \hat{\mathcal{L}}^\sigma}{\partial \boldsymbol{P}} + \boldsymbol{P}(t\text{-}1)$,
  $\pi(t) \leftarrow \epsilon \frac{\partial \hat{\mathcal{L}}^\sigma}{\partial \pi} + \pi(t\text{-}1)$
  $\boldsymbol{P}(t) \leftarrow$ precisions_clipping($\boldsymbol{P}$, $D_{\max}$) //see section 2.5
  $\pi(t) \leftarrow$ normalization($\pi(t)$) //see eq. (7)
  $\ell(t) \leftarrow (1-\alpha)\ell(t-1) + \alpha \hat{\mathcal{L}}^\sigma(\boldsymbol{x}(t))$ // sliding likelihood
  if *annealing update iteration* **then** // see section 2.4
    if $\Delta < \delta$ **then** // Δ see eq. (11)
      $\sigma(t) \leftarrow 0.9\sigma(t-1), \epsilon(t) \leftarrow 0.9\epsilon(t-1)$

Proposition The gradient $\frac{\partial \hat{\mathcal{L}}^\sigma}{\partial \sigma}$ is strictly positive for $\sigma > 0$
Proof For each sample, the 2D profile of $\log(\pi_k \mathcal{N}_k) \equiv f_k$ is assumed to be centered on the best-matching component k^* and depends on the distance from it as a function of $||k - k^*||$. We thus have $f_k = f_k(r)$ with $r \equiv ||k - k^*||$. Passing to the continuous domain, the indices in the Gaussian "smoothing filter" g_{k^*k} become continuous variables, and we have $g_{k^*k} \rightarrow g(||k - k^*||, \sigma) \equiv g(r, \sigma)$. Similarly, $f_k(r) \rightarrow f(r)$. Using 2D polar coordinates, the smoothed max-component likelihood $\hat{\mathcal{L}}^\sigma$ becomes a polar integral around the position of the best-matching component: $\hat{\mathcal{L}}^\sigma \sim \int_{\mathbb{R}^2} g(r,\sigma) f(r) dr d\phi$. It is trivial to show that for the special case of a constant log-probability profile, i.e., $f(r) = L$, $\mathcal{L}^\sigma$, does not depend on σ because Gaussians are normalized, and that the derivative w.r.t. σ vanishes:

$$\begin{aligned} \frac{d\hat{\mathcal{L}}^\sigma}{d\sigma} &\sim \int_0^\infty dr \Big(\frac{r^2}{\sigma^2} - 1\Big) \exp\Big(-\frac{r^2}{2\sigma^2}\Big) L \\ &= L \int_0^\sigma dr \Big(\frac{r^2}{\sigma^2} - 1\Big) \exp\Big(-\frac{r^2}{2\sigma^2}\Big) - L \int_\sigma^\infty \Big(\frac{r^2}{\sigma^2} - 1\Big) \exp\Big(-\frac{r^2}{2\sigma^2}\Big) \\ &\equiv L\mathcal{N} - L\mathcal{P} \end{aligned} \tag{12}$$

where we have split the integral into parts where the derivative w.r.t. σ is negative ($\mathcal{N}$) and positive ($\mathcal{P}$). We know that $\mathcal{N} = \mathcal{P}$ since the derivative must be zero for a constant function $f(r) = L$ due to the fact that Gaussians are normalized to the same value regardless of σ.

For a function $f(r)$ that satisfies $f(r) > L \forall r \in [0, \sigma[$ and $f(r) < L \forall r \in]\sigma, \infty[$, the inner and outer parts of the integral behave as follows:

$$\begin{aligned} \tilde{\mathcal{N}} &= \int_0^\sigma g(r) \Big(\frac{r^2}{\sigma^2} - 1\Big) f(r) < \int_0^\sigma g(r) \Big(\frac{r^2}{\sigma^2} - 1\Big) L = L\mathcal{N} \\ \tilde{\mathcal{P}} &= \int_\sigma^\infty g(r) \Big(\frac{r^2}{\sigma^2} - 1\Big) f(r) < \int_\sigma^\infty g(r) \Big(\frac{r^2}{\sigma^2} - 1\Big) L = L\mathcal{P} \end{aligned} \tag{13}$$

since $f(r)$ is minorized/majorized by L by assumption, and the contributions in both integrals have the same sign for the whole domain of integration. Thus, it is

shown that

$$\frac{d\hat{\mathcal{L}}^{\sigma}}{d\sigma} = \tilde{\mathcal{N}} - \tilde{\mathcal{P}} < L\mathcal{N} - L\mathcal{P} = 0 \tag{14}$$

for $\sigma > 0$ and, furthermore, that this derivative is zero for $\sigma = 0$ because $\hat{\mathcal{L}}^{\sigma}$ no longer depends on σ for this case.

Taking everything into consideration, in a situation where the log-likelihood $\hat{\mathcal{L}}^{\sigma}$ has reached a stationary point for a given value of σ, we have shown that:

- the value of $\hat{\mathcal{L}}^{\sigma}$ depends on σ,
- without changing the log-probabilities, we can increase $\hat{\mathcal{L}}^{\sigma}$ by reducing σ, assuming that the log-probabilities are mildly decreasing around the BMU,
- increasing $\hat{\mathcal{L}}^{\sigma}$ works as long as $\sigma > 0$. At $\sigma = 0$ the derivative becomes 0.

Thus, σ indeed defines an upper bound to $\hat{\mathcal{L}}^{\sigma}$ which can be increased by decreasing σ. The assumption of log-probabilities that decrease, on average, around the BMU is reasonable, since such a profile maximizes $\hat{\mathcal{L}}^{\sigma}$. All functions $f(r)$ that, e.g., decrease monotonically around the BMU, fulfill this criterion, whereas the form of the decrease is irrelevant.

2.7 Training Procedure for sEM

We use sEM proposed by [**?**] as a reference point to which we compare our SGD approach. We choose the step size of the form $\rho_t = \rho_0(t+1)^{-0.5+\alpha}$, with $\alpha \in [0, 0.5]$, $\rho_0 < 1$ and enforce $\rho(t) \geq \rho_\infty$. Values for these parameters are determined via a grid search in the ranges $\rho_0 \in \{0.01, 0.05, 0.1\}$, $\alpha \in \{0.01, 0.25, 0.5\}$ and $\rho_\infty \in \{0.01, 0.001, 0.0001\}$. Each sEM iteration uses a batch size B. Initial accumulation of sufficient statics is conducted for 10% of an epoch. Parameter initialization and clipping of precisions is performed just as for SGD, see section 2.5.

2.8 Comparing SGD and sEM

Since sEM optimizes the log-likelihood $\mathcal{L}$, whereas SGD optimizes the annealed approximation $\hat{\mathcal{L}}^{\sigma}$, the comparison of these measures should be considered carefully. We claim that the comparison is fair assuming that **i)** SGD annealing has converged and **ii)** GMM responsibilities are sharply peaked so that a single component has a responsibility of ≈ 1. It follows from **i)** that $\hat{\mathcal{L}}^{\sigma} \approx \hat{\mathcal{L}}$ and **ii)** implies that $\hat{\mathcal{L}} \approx \mathcal{L}$. Condition **ii)** is usually satisfied to high precision especially for high-dimensional inputs: if it is not, the comparison is biased in favor of sEM, since $\mathcal{L} > \hat{\mathcal{L}}$ by definition.

3 Experiments

Unless stated otherwise, the experiments in this section will be conducted with the following parameter values for sEM and SGD (where applicable): mini-batch size $B = 1$, $K = 8 \times 8$, $\mu^i = 0.1$, $\sigma_0 = 2$, $\sigma_\infty = 0.01$, $\epsilon = 0.001$, $D_{\max} = 20$. Each experiment is repeated 10 times with identical parameters but different random seeds for parameter initialization. See section 5 for a justification of these choices. Due

to input dimensionality, all precision matrices are assumed to be diagonal. The training/test data comes from the datasets shown below (see section 3.1).

3.1 Datasets

We use a variety of different image-based datasets, as well as a non-image dataset for evaluation purposes. All datasets are normalized to the $[0, 1]$ range.
MNIST [**?**] contains gray-scale images, which depict handwritten digits from 0 to 9 in a resolution of 28×28 pixels – the common benchmark for computer vision systems and classification problems.
SVHN [**?**] contains color images of house numbers (0-9, resolution 32×32).
FashionMNIST [**?**] contains grayscale images of 10 clothing categories and is considered as a more challenging classification task compared to MNIST.
Fruits 360 [**?**] consists of color pictures ($100 \times 100 \times 3$ pixels) showing different types of fruits. The ten best-represented classes are selected.
Devanagari [**?**] includes grayscale images of handwritten Devanagari letters with a resolution of 32×32 pixels – the first 10 classes are selected.
NotMNIST [**?**] is a grayscale image dataset (resolution 28×28 pixels) of letters from A to J extracted from different publicly available fonts.
ISOLET [**?**] is a non-image dataset containing 7 797 samples of spoken letters recorded from 150 subjects. Each sample is encoded and is represented by 617 float values.

3.2 Robustness of SGD to Initial Conditions

Here, we train GMM for three epochs on classes 1 to 9 for each dataset. We use different random and non-random initializations of the centroids and compare the final log-likelihood values. Random centroid initializations are parameterized by $\mu^i \in \{0.1, 0.3, 0.5\}$, whereas non-random initializations are defined by centroids from a previous training run on class 0 (one epoch). The latter is done to have a non-random centroid initialization that is as dissimilar as possible from the training data. The initialization of the precisions cannot be varied, because empirical data shows that training converges to undesirable solutions if the precisions are not initialized to large values. While this will have to be investigated further, we find that convergence to near-identical levels, regardless of centroid initialization for all datasets (see section 3.3 for more details).

3.3 Added Value of Annealing

To demonstrate the beneficial effects of annealing, we perform experiments on all datasets with annealing turned off. This is achieved by setting $\sigma_0 = \sigma_\infty$. This invariably produces sparse-component solutions with strongly inferior log-likelihoods after training, please refer to section 3.3.

Table 1 Effect of different random and non-random centroid initializations of SGD training. Given are the means and standard deviations of final log-likelihoods (10 repetitions per experiment). To show the added value of annealing, the right-most column indicates the final log-likelihoods when annealing is turned off. This value should be compared to the leftmost entry in each row where annealing is turned on. Standard deviations in this case where very small so they are omitted.

Initialization	**random**						**non-random**		no annealing
	$\mu^i = 0.1$		$\mu^i = 0.3$		$\mu^i = 0.5$		init class 0		$\mu^i = 0.1$
Dataset	*mean*	*std*	*mean*	*std*	*mean*	*std*	*mean*	*std*	*mean*
MNIST	205.47	1.08	205.46	0.77	205.68	0.78	205.37	0.68	124.1
FashionMNIST	231.22	1.53	231.58	2.84	231.00	1.11	229.59	0.59	183.0
NotMNIST	−48.41	1.77	−48.59	1.56	−48.32	1.13	−49.37	2.32	-203.8
Devanagari	−15.95	1.59	−15.76	1.34	−17.01	1.11	−22.07	4.59	-263.4
Fruits 360	12 095.80	98.02	12 000.70	127.00	12 036.25	122.06	10 912.79	1727.61	331.2
SVHN	1328.06	0.94	1327.99	1.59	1328.40	1.17	1327.80	0.94	863.2
ISOLET	354.34	0.04	354.36	0.04	354.36	0.04	354.20	0.05	201.5

3.4 Clustering Performance Evaluation

To compare the clustering performance of sEM and GMM the Davies-Bouldin score [**?**] and the Dunn index [**?**] are determined. We evaluate the grid-search results to find the best parameter setup for each metric for comparison. sEM is initialized by k-means to show that our approach does not depend on parameter initialization. Section 3.4 indicaties that SGD can egalize sEM performance.

Table 2 Clustering performance comparison of SGD and sEM training using Davies-Bouldin score (less is better) and Dunn index (more is better). Each time mean metric value (of 10 experiment repetitions) at the end of training, and their standard deviations are presented. Results are in bold face whenever they are better by more than half a standard deviation.

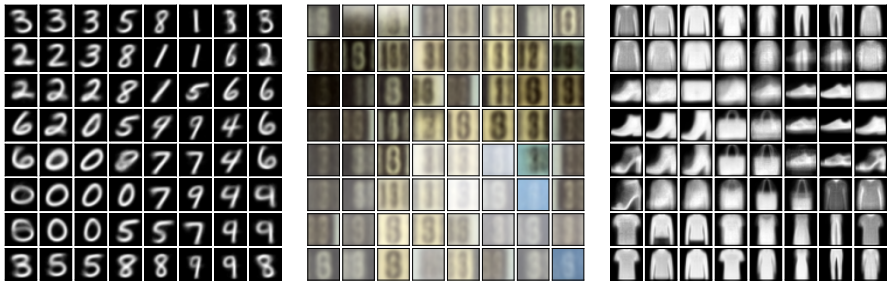
Metric	**Davies-Bouldin score**				**Dunn index**			
Algo.	**SGD**		**sEM**		**SGD**		**sEM**	
Dataset	*mean*	*std*	*mean*	*std*	*mean*	*std*	*mean*	*std*
MNIST	2.50	0.04	**2.47**	0.04	0.18	0.02	0.16	0.02
FashionMNIST	**2.06**	0.05	2.20	0.04	0.20	0.03	0.19	0.02
NotMNIST	2.30	0.03	**2.12**	0.03	0.15	0.03	0.14	0.04
Devanagari	**2.60**	0.04	2.64	0.02	**0.33**	0.01	0.27	0.04
SVHN	**2.34**	0.04	2.41	0.03	0.15	0.02	0.15	0.02

3.5 Streaming Experiments with Constant Statistics

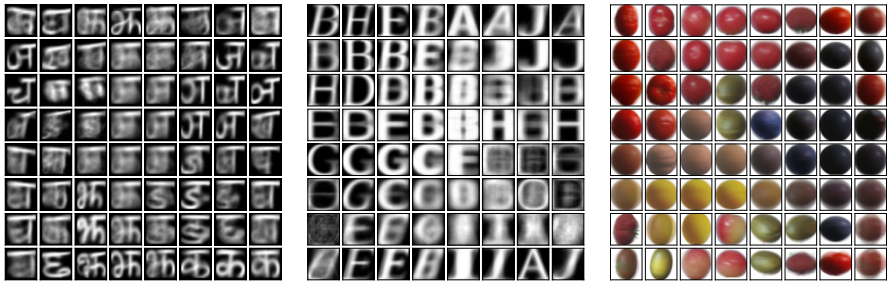
We train GMMs for three epochs (enough for convergence in all cases) using SGD and sEM on all datasets as described in sections 2.5 and 2.7. The resulting centroids of our SGD-based approach are shown in fig. 3, whereas the final loss values for SGD and sEM are compared in section 3.5. The centroids for both approaches are visually similar, except for the topological organization due to annealing for SGD, and the fact that in most experiments, some components do not converge for sEM while the others do. Section 3.5 indicates that SGD achieves performances superior to sEM in the majority of cases, in particular for the highest-dimensional datasets (SVHN: 3072 and Fruits 360: 30 000 dimensions).

Table 3 Comparison of SGD and sEM training on all datasets in a streaming-data scenario. Shown are log-likelihoods at the end of training, averaged over 10 repetitions, along with their standard deviations. Results are in bold face whenever they are higher by more than half a standard deviation. Additionally, the averaged maximum responsibilities (p_{k^*}) for test data are given for justifying the max-component approximation.

Algorithm / Dataset	**SGD**			**sEM**	
	$\varnothing \max p_{k^*}$	*mean*	*std*	*mean*	*std*
MNIST	0.992 674	216.6	0.31	216.8	1.38
FashionMNIST	0.997 609	**234.5**	2.28	222.9	6.03
NotMNIST	0.998 713	**−34.7**	1.16	−40.0	8.90
Devanagari	0.999 253	**−14.6**	1.09	−13.4	6.16
Fruits 360	0.999 746	**11754.3**	75.63	5483.0	1201.60
SVHN	0.998 148	**1329.8**	0.80	1176.0	16.91
ISOLET	0.994 069	354.2	0.01	354.5	0.37



(a) MNIST (b) SVHN (c) FashionMNIST



(d) Devanagari (e) NotMNIST (f) Fruits 360

Fig. 3 Exemplary results for centroids learned by SGD.

Visualization of High-dimensional sEM Outcomes Section 3.5 was obtained after training GMMs by sEM on both the Fruits 360 and the SVHN dataset. It should be compared to fig. 3, where an identical procedure was used to visualize centroids of SGD-trained GMMs. It is notable that the effect of unconverged components does not occur at all for our SGD approach, which is due to the annealing mechanism that "drags" unconverged components along.

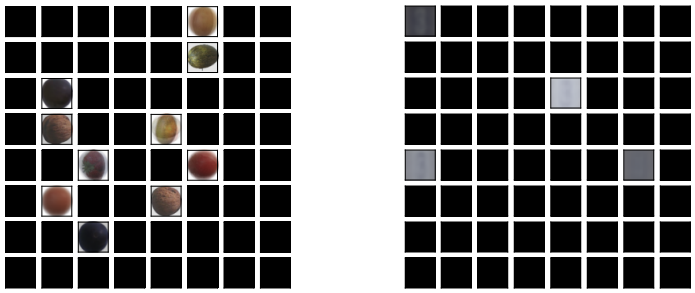


Fig. 4 Visualization of centroids after training runs (3 epochs) on high-dimensional datasets for sEM: Fruits 360 (left, 30 000 dimensions) and SVHN (right, 3000 dimensions). Component entries are displayed "as is", meaning that low brightness means low RGB values. Many GMM components remain unconverged, which is analogous to a sparse-component solution and explains the low log-likelihood values for these high-dimensional datasets.

4 Assumptions made by EM and SGD

The EM algorithm assumes that the observed data samples $\{\boldsymbol{x}_n\}$ depend on unobserved latent variables $\boldsymbol{z}_n$ in a non-trivial fashion, see section 2. The derivation of the EM algorithm starts out with the total incomplete-data log-likelihood

$$\begin{aligned}\mathcal{L} &= \log p(X) = \log \prod_n p(\boldsymbol{x}_n) = \sum_n \log p(\boldsymbol{x}_n) \\ &= \sum_n \log \sum_k p(\boldsymbol{x}_n, z_n = k) \\ &= \sum_n \log \sum_k p(z_n = k) \frac{p(\boldsymbol{x}_n, z_n = k)}{p(z_n = k)}.\end{aligned} \tag{15}$$

Due to the assumption that $\mathcal{L}$ is obtained by marginalizing out the latent variables, an explicit dependency on z_n can be re-introduced. For the last expression, Jensen' inequality can be used to construct a lower bound:

$$\begin{aligned}\mathcal{L} &\sim \sum_n \log \sum_k p(\boldsymbol{z}_n = k) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n = k)}{p(\boldsymbol{z}_n = k)} \\ &\geq \sum_n \sum_k p(\boldsymbol{z}_n = k) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n = k)}{p(\boldsymbol{z}_n = k)}.\end{aligned} \tag{16}$$

Since the realizations of the latent variables are unknown, we can assume any form for their distribution. In particular, for the choice $p(z_n) \sim p(\boldsymbol{x}_n, z_n)$, the lower bound becomes tight. Simple algebra and the fact that the distribution $p(z_n)$ must be normalized gives us:

$$\begin{aligned}p(z_n = k) &= \frac{p(z_n = k, \boldsymbol{x}_n)}{p(\boldsymbol{x}_n)} \\ &= p(z_n = k | \boldsymbol{x}_n) \\ &= \frac{p(z_n = k, \boldsymbol{x}_n)}{\sum_l p(z_n = l, \boldsymbol{x}_n)} \\ &= \frac{\pi_k \mathcal{N}_k(\boldsymbol{x}_n)}{\sum_l \pi_l \mathcal{N}_l(\boldsymbol{x}_n)}\end{aligned} \tag{17}$$

where we have used eq. (15) in the last step. $p(z_n = k|\boldsymbol{x}_n)$ is a quantity that can be computed from data with no reference to the latent variables. For GMM it is usually termed *responsibility* and we write it as $p(z_n = k|\boldsymbol{x}_n) \equiv \gamma_{nk}$.

However, the construction of a tight lower bound, which is actually different from $\mathcal{L}$, only works when $p(\boldsymbol{x}_n, z_n)$ depends non-trivially on the latent variable z_n. If this is not the case, we have $p(\boldsymbol{x}_n, z_n) = K^{-1} p(\boldsymbol{x}_n)$ and the derivation of eq. (17) goes down very differently:

$$\begin{aligned}
\mathcal{L} \sim \sum_n \log p(\boldsymbol{x}_n) &\geq \sum_n \sum_k p(z_n = k) \log \frac{p(\boldsymbol{x}_n, z_n = k)}{p(z_n = k)} \\
&= \sum_n \sum_k p(z_n = k) \log \frac{K^{-1} p(\boldsymbol{x}_n)}{p(z_n = k)} \\
&= \sum_n \log \Big(K^{-1} p(\boldsymbol{x}_n) \Big) - \sum_k p(z_n = k) \log p(z_n = k) \\
&\equiv \sum_n \Big(\log p(\boldsymbol{x}_n) - \big(\log K - \mathcal{H}[z_n] \big) \Big)
\end{aligned} \tag{18}$$

where $\mathcal{H}$ represents the Shannon entropy of $p(\boldsymbol{z})$. The highest value this can have is $\log K$ for an uniform distribution of the z_n, finally leading to a lower bound for $\mathcal{L}$ of

$$\mathcal{L} \geq \sum_n \Big(\log p(\boldsymbol{x}_n) \Big) \tag{19}$$

which is trivial by Jensen's inequality, but not tight. In particular, no closed-form solutions to the associated extremal value problem can be computed.

This shows that optimizing GMM by EM assumes that each sample has been drawn from a single element in a set of K uni-modal Gaussian distributions. Which distribution is selected for sampling depends on a latent random variable. On the other hand, optimization by SGD uses the incomplete-data log-likelihood $\mathcal{L}$ as basis for optimization, without assuming the existence of hidden variables at all. This may be advantageous for problems where the assumption of Gaussianity is badly violated, although empirical studies indicate that optimization by EM works very well in a very wide range of scenarios.

5 Discussion and Conclusion

The **relevance of this article** is outlined by the fact that training GMMs by SGD was recently investigated in the community by [**?**,**?**]. We go beyond, since our approach does not rely on off-line data-driven model initialization, and works for high-dimensional streaming data. The presented SGD scheme is simple and very robust to initial conditions due to the proposed annealing procedure, see section 3.2 and section 3.3. In addition, our SGD approach compares favorably to the reference model for online EM [**?**] in terms of achieved log-likelihoods, which was verified on multiple real-world datasets. Superior SGD performance is observed for the high-dimensional datasets.

Analysis of results suggests that SGD performs better than sEM on average, see section 3.5, although the differences are very modest. It should be stated clearly that it cannot be expected, and is not the goal of this article, to outperform sEM

by SGD in the general case, only to achieve a similar performance. However, if sEM is used without, e.g., k-means initialization, components may not converge (see section 3.5) for very high-dimensional data like Fruits 360 and SVHN datasets, which is why SGD outperforms sEM in this case. Another important advantage of SGD over sEM is the fact that the only parameter that needs to be tuned is the learning rate ϵ, whereas sEM has a complex and not intuitive dependency on ρ_0, ρ_∞ and α_0.

Small batch sizes and streaming data are possible with the SGD-based approach. Throughout the experiments, we used a batch size of 1, which allows streaming-data processing without the need to store any samples at all. Larger batch sizes are possible and strongly increase execution speed. In the conducted experiments, SGD (and sEM) usually converged within the first two epochs, which is a substantial advantage whenever huge sets of data have to be processed.

Low-dimensional data can be treated with our SGD-based approach, either using the max-component approximation of eq. (8) or the full incomplete-data log-likelihood. For low-dimensional data, numerical issues and undesirable local minima are less relevant, so there really is no point using the max-component approximation together with annealing here. Extensive experiments with synthetic data drawn from various Gaussian mixture distributions show that this is nevertheless possible, with parameters identical to the experiments inn section 3, if the initialization range of the centroids, μ^i, is chosen to coincide with the ranges of the individual data components. This ensures that initial centroids cover the data space sufficiently so that each cluster in the data has at least one centroid near to it.

No assumptions about data generation are made by SGD in contrast to the EM and sEM algorithms. The latter guarantees that the loss will not decrease due to an M-step. This, however, assumes a non-trivial dependency of the data on an unobservable latent variable (shown in section 4). In contrast, SGD makes no hard-to-verify assumptions, which is a rather philosophical point, but may be an advantage in certain situations where data are strongly non-Gaussian.

Numerical stability is assured by our SGD training approach. It does not optimize the log-likelihood but its max-component approximation. This approximation contains no exponentials at all, and is well justified by the results of section 3.5 which shows that component probabilities are strongly peaked. In fact, it is the gradient computations where numerical problems occurred, e.g., NaN values. The "logsumexp" trick mitigates the problem, but does not eliminate it (see section 2.2). It cannot be used when gradients are computed automatically, which is what most machine learning frameworks do.

Hyper-Parameter selection guidelines are as follows: the learning rate ϵ must be set by cross-validation (a good value is 0.001). We empirically found that initializing precisions to the cut-off value $D_{\max}$ and an uniform initialization of the π_i are beneficial, and that centroids are best initialized to small random values. A value of $D_{\max} = 20$ always worked in our experiments. Generally, the cut-off must be much larger than the inverse of the data variance. In many cases, it should be possible to estimate this roughly, even in streaming settings, especially when samples are normalized. For density estimation, choosing higher values for K leads to higher final log-likelihoods. For clustering, K should be selected using standard techniques for GMMs. The parameter δ controls loss stationarity detection for the

annealing procedure and was shown to perform well for $\delta = 0.05$. Larger values will lead to a faster decrease of $\sigma(t)$, which may impair convergence. Smaller values are always admissible but lead to longer convergence times. The annealing time constant α should be set to the GMM learning rate ϵ or lower. Smaller values of α lead to longer convergence times since $\sigma(t)$ will be updated less often. The initial value σ_0 needs to be large in order to enforce convergence for all components. A typical value is $\sqrt{K}$. The lower bound on σ_∞ should be as small as possible in order to achieve high log-likelihoods (e.g., 0.01, see section 2.6 for a proof).

6 Future Work

The presented work can be extended in several ways: First of all, annealing control could be simplified further by inferring good δ values from α. Likewise, *increases* of σ might be performed automatically when the loss rises sharply, indicating a task boundary. As we found that GMM convergence times grow linear with the number of components, we will investigate hierarchical GMM models that operate like a Convolutional Neural Network (CNN), in which individual GMM only see a local patch of the input and can therefore have low K.

Conflict of Interest

The authors declare that they have no conflict of interest.