Pattern Recognition Letters
journal homepage: www.elsevier.com

Hybrid generative-discriminative training of Gaussian mixture models: Supplementary material

Wolfgang Roth[a,**], Robert Peharz[b], Sebastian Tschiatschek[c], Franz Pernkopf[a]

[a] *Graz University of Technology, Inffeldgasse 16c/EG, 8010 Graz, Austria*
[b] *University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom*
[c] *ETH Zürich, Universitätstrasse 6, 8092 Zürich, Switzerland*

1. Additional information about the data sets

To obtain more challenging data sets, the MNIST data set has been transformed by various operations (Larochelle et al., 2007). In particular, there are the following variants:

- *MNIST Basic*: This data set has not been transformed. The data set is merely split differently into training, validation, and test set, respectively.
- *MNIST Background*: The background pixels of the images are replaced by random image patches.
- *MNIST Background Random*: The background pixels of the images are set to uniformly distributed random pixel values.
- *MNIST Rotated*: The images are randomly rotated.
- *MNIST Rotated Background*: The transformations used for MNIST Rotated and MNIST Background are combined

Some samples of the variants of MNIST and CIFAR-10 (Krizhevsky, 2009) are shown in Figure 1.

2. Experiments on synthetic data

2.1. Illustration of hyperparameters

The synthetic example shown in Figure 2, illustrates the influence of λ and γ on different aspects of hybrid LM GMMs. The data comprises samples of two classes which are shown as red and blue points respectively. A GMM with $K_{c_{red}} = K_{c_{blue}} = 2$ was learned for each class by generative training with the EM algorithm (Dempster et al., 1977) followed by hybrid training with the ADAM algorithm (Kingma and Ba, 2015). The first row shows the decision boundary (green line) and the lines where the desired log-margin γ is satisfied with equality. Red points falling below the red line and blue points falling above the blue line violate the desired log-margin and are therefore penalized in the large margin term of the hybrid objective. The second row shows the class posterior probabilities of the red points $p(c_{red}|\boldsymbol{x}, \boldsymbol{\theta})$. The last row shows the contour lines of the logarithm of the class conditional distributions $p(\boldsymbol{x}|\boldsymbol{\theta}_c)$ for both classes.

The first three columns show how the model changes for varying λ and fixed $\gamma = 1$. In Figure 2(a), the ML model is unaware of the samples close to the decision boundary. The behavior slightly changes in Figure 2(b) when we decrease λ: The decision boundary gets aligned at the points so that several of them satisfy the desired log-margin with equality. By further increasing the discriminative character of the model in Figure 2(c), all data points are classified correctly. The corresponding class posterior probabilities in Figures 2(f), (g), and (h) show that the uncertainty close to the decision boundary is reasonable. Figures 2(k), (l), and (m) show how the data fit degrades slightly as the discriminative semantics is increased.

Nevertheless, the data is captured at a decent level in all three parameter settings since the relatively small desired log-margin parameter $\gamma = 1$ is achievable by all points by only slightly modifying the ML parameters. In this case, the hybrid objective is able to select the model with the best data fit among those models where all samples satisfy the desired log-margin. The situation changes as we increase the desired log-margin parameter γ. In Figures 2(d) and (e), the region between both desired-margin lines has grown and many samples violate the desired log-margin. In this case, many samples are considered in the LM term, and, therefore, individual samples have less effect on the decision boundary. However, as shown in Figures 2(i) and (j), the probabilistic semantic is largely abandoned as is indicated by the almost instant change between zero and one when moving only slightly away from the decision boundary. Also the contour lines of the log-pdfs in Figures 2(n) and (o) show that the generative semantics is abandoned. The means tend to

** Corresponding author
e-mail: roth@tugraz.at (Wolfgang Roth)

(a)

(b)

(c)

(d)

(e)

(f)

Fig. 1. Some samples from the MNIST data set, variants of the MNIST data set, and the CIFAR-10 data set. (a) MNIST and MNIST Basic, (b) MNIST Background, (c) MNIST Background Random, (d) MNIST Rotated, (e) MNIST Rotated Background, and (f) CIFAR-10.

move away from the data and the covariances shrink such that the desired log-margin is achieved by many samples.

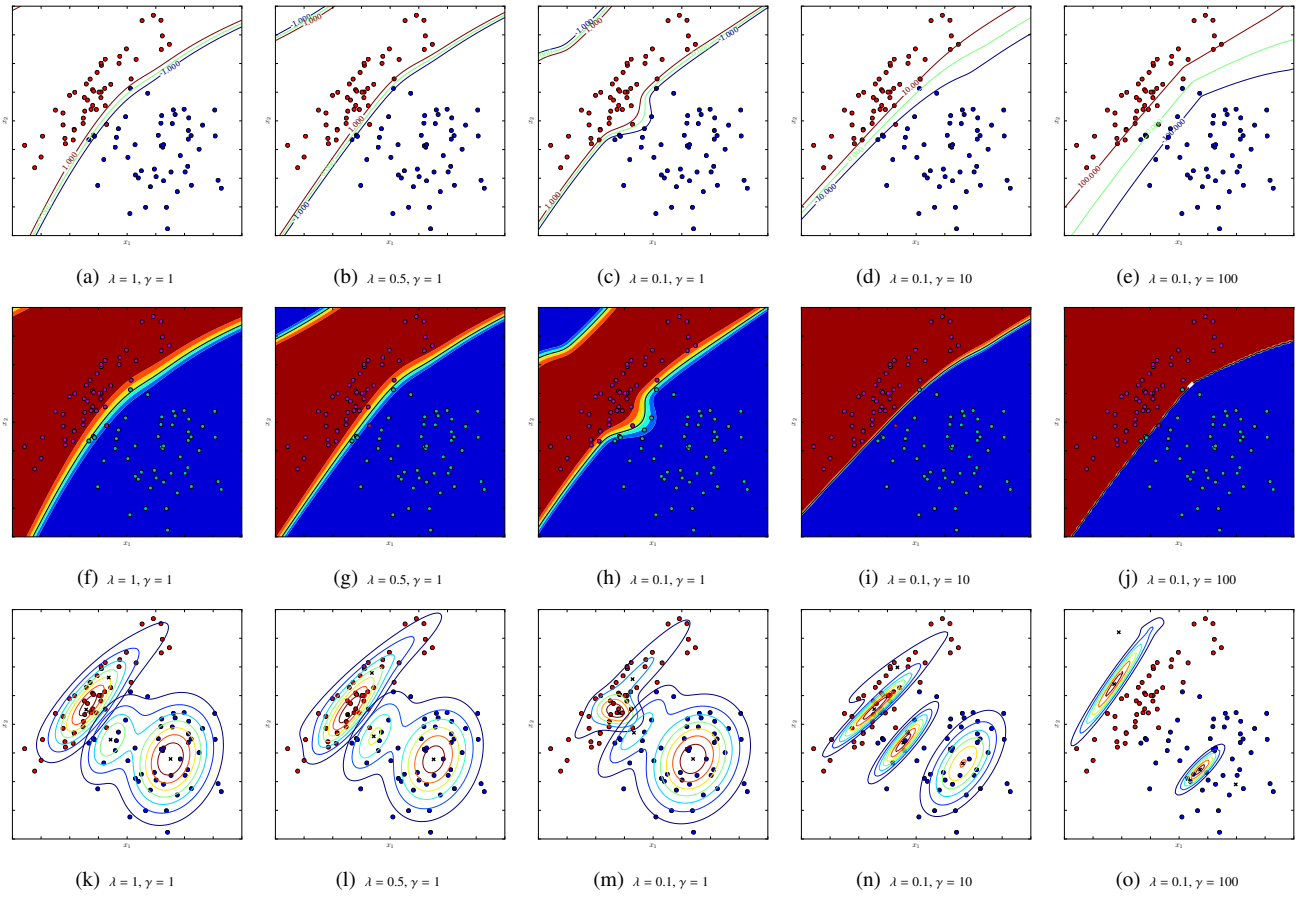
2.2. Spiral data set

The next experiment highlights some differences between diagonal and full covariance matrices. Therefore, we generated a synthetic data set containing two intertwined spirals each containing 500 samples. We trained GMMs with $K_{c_{red}} = K_{c_{blue}} = 7$ using the same procedure as in Section 2.1.

This is shown in Figure 3. The first row shows the class posterior probability of the red points $p(c_{red}|\boldsymbol{x}, \boldsymbol{\theta})$. The second and third rows show the class conditional densities $p(\boldsymbol{x}|c_{red}, \boldsymbol{\theta})$ and $p(\boldsymbol{x}|c_{blue}, \boldsymbol{\theta})$, respectively.

The ML solution for diagonal covariance matrices, shown in the first column, achieves only a large error. Since the diagonal covariance matrices are too weak to model the curved shape of the data distribution, the decision regions have a rather rectilinear shape, and the circular shape of the spiral is not reflected in the decision regions. Nevertheless, the model with diagonal covariance matrices in the second column is able to model the spiral shape in the decision regions by incorporating the discriminative LM term. This decreases the classification error substantially but comes at the cost of partly losing the generative semantics. This can be seen in the class conditional densities shown in Figure 3(f) and Figure 3(j). The Gaussian components are chosen such that their interaction achieves a spiral shaped decision boundary rather than to capture the data well. This can also be seen in the green region to the right of the center in Figure 3(b). Here, the model classifies the red points correctly, but the class posterior probability $p(c_{red}|\boldsymbol{x}, \boldsymbol{\theta})$ is only slightly larger than $p(c_{blue}|\boldsymbol{x}, \boldsymbol{\theta})$ although this area contains almost exclusively red points.

The behavior is better in case of full covariance matrices as shown in the third column. The ML solution captures the curved shape of the data distribution much better which trans-



(a) $\lambda = 1, \gamma = 1$ (b) $\lambda = 0.5, \gamma = 1$ (c) $\lambda = 0.1, \gamma = 1$ (d) $\lambda = 0.1, \gamma = 10$ (e) $\lambda = 0.1, \gamma = 100$

(f) $\lambda = 1, \gamma = 1$ (g) $\lambda = 0.5, \gamma = 1$ (h) $\lambda = 0.1, \gamma = 1$ (i) $\lambda = 0.1, \gamma = 10$ (j) $\lambda = 0.1, \gamma = 100$

(k) $\lambda = 1, \gamma = 1$ (l) $\lambda = 0.5, \gamma = 1$ (m) $\lambda = 0.1, \gamma = 1$ (n) $\lambda = 0.1, \gamma = 10$ (o) $\lambda = 0.1, \gamma = 100$

Fig. 2. Effect of λ and γ on a synthetic classification example with two classes. (a)-(e) Decision boundaries and log-margin line at γ. (f)-(j) Posterior probability of the red points $p(c_{red}|\boldsymbol{x}, \theta)$. We have $p(c_{red}|\boldsymbol{x}, \theta) \approx 1 \Rightarrow$ red, and $p(c_{red}|\boldsymbol{x}, \theta) \approx 0 \Rightarrow$ blue. (k)-(o) Contour lines of the logarithm of the class conditional distribution $p(\boldsymbol{x}|\theta_c)$ for both classes. The component means are shown as black crosses.

lates directly into a curved shape of the decision boundary. By incorporating the discriminative LM term as shown in the fourth column, the model is able to substantially decrease the classification error without degrading the generative semantics as severe as with diagonal covariance matrices.

3. Best hyperparameters

Tables 1–10 list the hyperparameters (obtained with Bayesian optimization) corresponding to the classification results of Table 1 in the main paper. We report the average number of components per class $\mu(K_c)$, the standard deviation of components per class $\sigma(K_c)$, the log of the regularizer for the diagonal of the covariance $\log_{10} \varepsilon$, the rank of the low-rank covariance approximation R, the hybrid generative-discriminative trade-off parameter λ, the log of the desired log-margin $\log_{10} \gamma$, and the log of the step size $\log_{10} \eta$. For the hybrid and discriminative objectives, we also report the number of training epochs needed to achieve the best validation performance for the corresponding hyperparameters. Note that the number of epochs differs for the individual data sets. For MNIST, CIFAR-10, and TIMIT (with the corresponding pca50 variants) we trained for 100 epochs. For the remaining data sets, we trained for 500 epochs.

Tables 11–14 list the hyperparameters corresponding to the semi-supervised classification results of Table 2 in the main paper. For these experiments, the additional semi-supervised trade-off parameter κ is reported. Furthermore, only a single shared number of components per class K is reported.

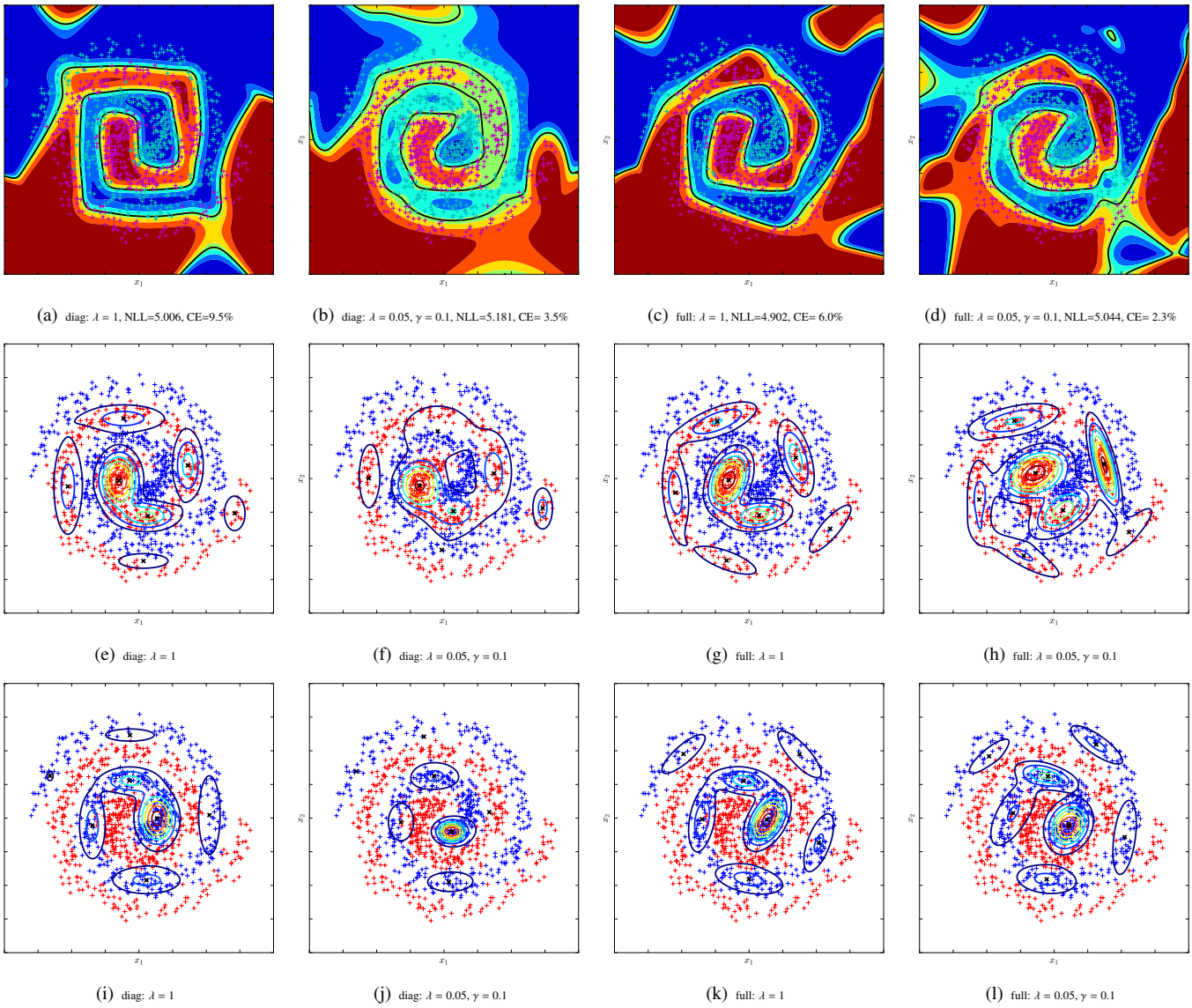
References

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B 39, 1–38.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: ICLR. ArXiv: 1412.6980.

Krizhevsky, A., 2009. Learning multiple layers of features from tiny images. Technical Report. University of Toronto.

Larochelle, H., Erhan, D., Courville, A.C., Bergstra, J., Bengio, Y., 2007. An empirical evaluation of deep architectures on problems with many factors of variation, in: ICML, pp. 473–480.



(a) diag: $\lambda = 1$, NLL=5.006, CE=9.5%
(b) diag: $\lambda = 0.05$, $\gamma = 0.1$, NLL=5.181, CE= 3.5%
(c) full: $\lambda = 1$, NLL=4.902, CE= 6.0%
(d) full: $\lambda = 0.05$, $\gamma = 0.1$, NLL=5.044, CE= 2.3%
(e) diag: $\lambda = 1$
(f) diag: $\lambda = 0.05$, $\gamma = 0.1$
(g) full: $\lambda = 1$
(h) full: $\lambda = 0.05$, $\gamma = 0.1$
(i) diag: $\lambda = 1$
(j) diag: $\lambda = 0.05$, $\gamma = 0.1$
(k) full: $\lambda = 1$
(l) full: $\lambda = 0.05$, $\gamma = 0.1$

Fig. 3. ML GMMs ($\lambda = 1$) and hybrid LM GMMs for both diagonal and full covariance matrices. (a)-(d) Posterior probability of the red points $p(c_{red}|\boldsymbol{x}, \boldsymbol{\theta})$. We have $p(c_{red}|\boldsymbol{x}, \boldsymbol{\theta}) \approx 1 \Rightarrow$ red, and $p(c_{red}|\boldsymbol{x}, \boldsymbol{\theta}) \approx 0 \Rightarrow$ blue. The classification errors (CE) and the negative log-likelihood per sample (NLL) are shown in the captions. (e)-(h) Contour lines of the logarithm of the class conditional distribution $p(\boldsymbol{x}|c_{red}, \boldsymbol{\theta})$. The component means are shown as black crosses. (i)-(l) The same contour lines for $p(\boldsymbol{x}|c_{blue}, \boldsymbol{\theta})$.

Table 1. Best hyperparameters for the generative objective using full covariance matrices (LL).

dataset	$\mu(K_c)$	$\sigma(K_c)$	$\log_{10} \varepsilon$
MNIST	17.0	3.8	-1
MNIST (pca50)	13.6	5.7	-1
MNIST Basic	16.2	4.9	-1
MNIST Basic (pca50)	9.5	5.6	-1
MNIST Background	1.0	0.0	-2
MNIST Background (pca50)	1.1	0.3	-4
MNIST Background Random	19.5	0.7	-1
MNIST Background Random (pca50)	2.2	1.0	-2
MNIST Rotated	15.2	4.9	-1
MNIST Rotated (pca50)	10.9	3.2	-1
MNIST Rotated Background	1.0	0.0	-1
MNIST Rotated Background (pca50)	3.5	1.7	-1
CIFAR-10	4.6	1.7	-2
CIFAR-10 (pca50)	7.0	2.1	-1
TIMIT	7.3	5.6	-1

Table 2. Best hyperparameters for the hybrid LM objective using DPLR covariance matrices (LL+LM).

dataset	$\mu(K_c)$	$\sigma(K_c)$	R	$\log_{10}\varepsilon$	λ	$\log_{10}\gamma$	$\log_{10}\eta$	best epoch
MNIST	17.0	3.8	25	-1	0.0822	2.000	-3.246	91
MNIST (pca50)	13.2	4.7	25	-3	0.2458	1.776	-2.683	35
MNIST Basic	16.2	4.9	25	-1	0.7975	2.000	-3.717	174
MNIST Basic (pca50)	9.5	5.6	24	-1	0.0089	1.484	-4.000	50
MNIST Background	9.9	0.9	25	-4	0.0042	2.000	-2.000	152
MNIST Background (pca50)	1.4	0.7	25	-2	0.3160	1.712	-4.002	439
MNIST Background Random	9.4	0.5	25	-3	0.0042	-2.000	-2.000	497
MNIST Background Random (pca50)	2.2	1.0	25	-2	0.9991	-2.000	-2.981	471
MNIST Rotated	15.2	4.9	25	-1	0.4611	1.746	-3.027	126
MNIST Rotated (pca50)	7.6	1.0	25	-2	0.7128	2.000	-2.113	417
MNIST Rotated Background	1.0	0.0	24	-1	0.5236	1.276	-3.341	186
MNIST Rotated Background (pca50)	1.1	0.3	25	-2	0.5658	1.722	-2.318	447
CIFAR-10	6.6	2.4	25	-1	0.0544	2.000	-3.845	34
CIFAR-10 (pca50)	6.8	4.7	25	-2	0.4794	2.000	-2.623	76
TIMIT	3.3	3.5	16	-4	0.1829	2.000	-2.709	93

Table 3. Best hyperparameters for the hybrid CLL objective using DPLR covariance matrices (LL+CLL).

dataset	$\mu(K_c)$	$\sigma(K_c)$	R	$\log_{10}\varepsilon$	λ	$\log_{10}\eta$	best epoch
MNIST	17.0	3.8	25	-1	0.0580	-3.368	25
MNIST (pca50)	13.6	5.7	22	-1	0.0009	-3.890	46
MNIST Basic	16.2	4.9	25	-1	0.9957	-3.092	293
MNIST Basic (pca50)	3.7	3.2	25	-3	0.0239	-2.094	291
MNIST Background	10.4	1.2	25	-3	0.0009	-2.000	383
MNIST Background (pca50)	1.1	0.3	24	-3	0.2807	-2.626	96
MNIST Background Random	10.6	0.8	19	-2	0.0025	-2.000	406
MNIST Background Random (pca50)	2.2	1.0	25	-2	0.9002	-2.444	137
MNIST Rotated	15.2	4.9	25	-1	0.2437	-2.960	253
MNIST Rotated (pca50)	10.9	3.2	24	-1	0.8923	-2.190	200
MNIST Rotated Background	1.0	0.0	22	-1	0.1166	-4.350	115
MNIST Rotated Background (pca50)	1.2	0.4	25	-4	0.7012	-2.291	402
CIFAR-10	4.6	1.7	25	-2	0.0012	-2.480	85
CIFAR-10 (pca50)	6.8	4.7	25	-2	0.6499	-2.729	92
TIMIT	3.3	3.5	1	-4	0.0011	-2.000	96

Table 4. Best hyperparameters for the discriminative LM objective using DPLR covariance matrices (LM).

dataset	$\mu(K_c)$	$\sigma(K_c)$	R	$\log_{10}\varepsilon$	$\log_{10}\gamma$	$\log_{10}\eta$	best epoch
MNIST	17.0	3.8	21	-1	1.675	-4.137	25
MNIST (pca50)	12.2	3.6	25	-2	1.644	-3.142	65
MNIST Basic	16.2	4.9	25	-1	2.000	-5.000	37
MNIST Basic (pca50)	9.5	5.6	24	-1	1.506	-4.951	375
MNIST Background	10.4	1.2	25	-3	1.998	-2.000	468
MNIST Background (pca50)	1.4	0.7	12	-2	1.430	-2.932	21
MNIST Background Random	9.4	0.5	25	-3	2.000	-2.000	318
MNIST Background Random (pca50)	2.1	1.0	25	-4	1.319	-3.264	8
MNIST Rotated	15.2	4.9	21	-1	1.850	-3.669	138
MNIST Rotated (pca50)	5.2	2.0	25	-4	1.674	-4.069	116
MNIST Rotated Background	1.0	0.0	23	-1	1.236	-4.961	104
MNIST Rotated Background (pca50)	3.5	1.7	5	-1	1.153	-3.636	166
CIFAR-10	6.6	2.4	25	-1	2.000	-4.110	20
CIFAR-10 (pca50)	6.8	4.7	16	-2	1.603	-3.685	91
TIMIT	3.3	3.5	14	-4	1.989	-3.275	43

Table 5. Best hyperparameters for the discriminative CLL objective using DPLR covariance matrices (CLL).

dataset	$\mu(K_c)$	$\sigma(K_c)$	R	$\log_{10}\varepsilon$	$\log_{10}\eta$	best epoch
MNIST	17.0	3.8	25	-1	-3.644	72
MNIST (pca50)	13.6	5.7	25	-1	-3.335	50
MNIST Basic	16.2	4.9	25	-1	-3.890	1
MNIST Basic (pca50)	9.5	5.6	23	-1	-3.465	309
MNIST Background	1.0	0.0	25	-2	-2.000	270
MNIST Background (pca50)	1.6	1.0	12	-1	-3.882	24
MNIST Background Random	10.6	0.8	8	-2	-3.038	1
MNIST Background Random (pca50)	2.2	1.0	7	-2	-4.196	202
MNIST Rotated	15.2	4.9	25	-1	-2.909	82
MNIST Rotated (pca50)	10.9	3.2	25	-1	-4.878	484
MNIST Rotated Background	1.0	0.0	22	-1	-5.000	108
MNIST Rotated Background (pca50)	3.5	1.7	2	-1	-3.468	86
CIFAR-10	4.6	1.7	25	-2	-2.353	88
CIFAR-10 (pca50)	5.8	3.8	4	-4	-2.000	16
TIMIT	3.3	3.5	1	-4	-2.000	89

Table 6. Best hyperparameters for the generative objective using diagonal covariance matrices (LL).

dataset	$\mu(K_c)$	$\sigma(K_c)$	$\log_{10}\varepsilon$
MNIST	77.2	29.5	-1
MNIST (pca50)	88.2	15.2	-1
MNIST Basic	76.1	28.4	-1
MNIST Basic (pca50)	69.2	19.9	-1
MNIST Background	1.0	0.0	-4
MNIST Background (pca50)	10.4	4.3	-1
MNIST Background Random	65.5	24.0	-1
MNIST Background Random (pca50)	63.8	30.4	-1
MNIST Rotated	92.8	17.1	-1
MNIST Rotated (pca50)	81.2	20.3	-1
MNIST Rotated Background	92.8	13.2	-1
MNIST Rotated Background (pca50)	14.9	5.6	-3
CIFAR-10	95.9	7.3	-2
CIFAR-10 (pca50)	73.7	24.7	-1
TIMIT	62.1	28.6	-2

Table 7. Best hyperparameters for the hybrid LM objective using diagonal covariance matrices (LL+LM).

dataset	$\mu(K_c)$	$\sigma(K_c)$	$\log_{10}\varepsilon$	λ	$\log_{10}\gamma$	$\log_{10}\eta$	best epoch
MNIST	77.2	29.5	-1	0.0069	1.991	-2.469	86
MNIST (pca50)	91.7	14.8	-4	0.0546	1.999	-2.000	83
MNIST Basic	76.1	28.4	-1	0.0011	2.000	-2.000	142
MNIST Basic (pca50)	69.2	19.9	-1	0.0009	1.904	-2.001	28
MNIST Background	83.5	25.6	-1	0.0009	2.000	-2.000	261
MNIST Background (pca50)	10.4	4.3	-1	0.0422	2.000	-2.000	38
MNIST Background Random	65.5	24.0	-1	0.0009	1.317	-4.242	312
MNIST Background Random (pca50)	45.9	24.2	-3	0.0009	2.000	-2.404	79
MNIST Rotated	92.8	17.1	-1	0.0009	2.000	-2.553	13
MNIST Rotated (pca50)	80.0	27.9	-3	0.1604	2.000	-2.000	46
MNIST Rotated Background	92.8	13.2	-1	0.0009	2.000	-2.000	473
MNIST Rotated Background (pca50)	24.6	7.3	-1	0.0010	2.000	-2.890	445
CIFAR-10	61.6	35.3	-1	0.0009	2.000	-2.160	78
CIFAR-10 (pca50)	62.8	34.1	-3	0.0330	2.000	-2.442	99
TIMIT	60.9	28.8	-4	0.0082	2.000	-2.778	28

Table 8. Best hyperparameters for the hybrid CLL objective using diagonal covariance matrices (LL+CLL).

dataset	$\mu(K_c)$	$\sigma(K_c)$	$\log_{10}\varepsilon$	λ	$\log_{10}\eta$	best epoch
MNIST	77.2	29.5	-1	0.0009	-2.517	20
MNIST (pca50)	88.2	15.2	-1	0.0009	-2.000	50
MNIST Basic	76.1	28.4	-1	0.0009	-3.047	10
MNIST Basic (pca50)	42.4	21.3	-3	0.0009	-2.000	108
MNIST Background	1.0	0.0	-4	0.1482	-2.975	168
MNIST Background (pca50)	9.9	3.6	-3	0.1227	-2.041	4
MNIST Background Random	65.5	24.0	-1	0.0019	-5.000	103
MNIST Background Random (pca50)	45.9	24.2	-3	0.0009	-2.000	97
MNIST Rotated	81.7	20.0	-2	0.0041	-2.156	362
MNIST Rotated (pca50)	81.2	20.3	-1	0.0019	-2.009	497
MNIST Rotated Background	1.0	0.0	-4	0.2194	-3.344	254
MNIST Rotated Background (pca50)	24.6	7.3	-1	0.2045	-2.000	115
CIFAR-10	95.9	7.3	-2	0.0010	-2.534	90
CIFAR-10 (pca50)	73.7	24.7	-1	0.2656	-2.410	2
TIMIT	76.5	27.6	-1	0.2974	-2.465	1

Table 9. Best hyperparameters for the discriminative LM objective using diagonal covariance matrices (LM).

dataset	$\mu(K_c)$	$\sigma(K_c)$	$\log_{10}\varepsilon$	$\log_{10}\gamma$	$\log_{10}\eta$	best epoch
MNIST	77.2	29.5	-1	1.880	-2.789	64
MNIST (pca50)	81.8	18.4	-3	2.000	-2.311	59
MNIST Basic	76.1	28.4	-1	2.000	-2.580	225
MNIST Basic (pca50)	69.2	19.9	-1	2.000	-2.000	18
MNIST Background	83.5	25.6	-1	2.000	-2.006	331
MNIST Background (pca50)	10.0	5.0	-4	2.000	-2.874	202
MNIST Background Random	65.5	24.0	-1	1.694	-3.669	131
MNIST Background Random (pca50)	45.9	24.2	-3	2.000	-2.256	17
MNIST Rotated	92.8	17.1	-1	2.000	-3.600	107
MNIST Rotated (pca50)	79.8	20.4	-4	1.816	-2.255	21
MNIST Rotated Background	92.8	13.2	-1	2.000	-2.000	414
MNIST Rotated Background (pca50)	24.6	7.3	-1	2.000	-2.144	116
CIFAR-10	61.6	35.3	-1	2.000	-2.174	93
CIFAR-10 (pca50)	62.8	34.1	-3	1.669	-2.768	81
TIMIT	60.9	28.8	-4	1.925	-3.156	63

Table 10. Best hyperparameters for the discriminative CLL objective using diagonal covariance matrices (CLL).

dataset	$\mu(K_c)$	$\sigma(K_c)$	$\log_{10}\varepsilon$	$\log_{10}\eta$	best epoch
MNIST	77.2	29.5	-1	-2.698	67
MNIST (pca50)	88.2	15.2	-1	-2.402	84
MNIST Basic	76.1	28.4	-1	-2.485	200
MNIST Basic (pca50)	69.2	19.9	-1	-2.000	321
MNIST Background	1.0	0.0	-2	-3.309	10
MNIST Background (pca50)	10.0	5.0	-4	-2.000	3
MNIST Background Random	65.5	24.0	-1	-3.896	235
MNIST Background Random (pca50)	45.9	24.2	-3	-2.000	53
MNIST Rotated	81.7	20.0	-2	-2.252	498
MNIST Rotated (pca50)	81.2	20.3	-1	-2.970	222
MNIST Rotated Background	92.8	13.2	-1	-2.146	300
MNIST Rotated Background (pca50)	14.9	5.6	-3	-2.008	1
CIFAR-10	95.9	7.3	-2	-2.536	78
CIFAR-10 (pca50)	73.7	24.7	-1	-2.372	1
TIMIT	76.5	27.6	-1	-2.606	1

Table 11. Best hyperparameters of the semi-supervised experiment on MNIST (pca50) for the supervised hybrid LM objective (SV LM).

N_l	K	$\log_{10}\varepsilon$	R	λ	$\log_{10}\gamma$	$\log_{10}\eta$	best epoch
100	1	-1	13	0.9991	2.000	-2.000	48
250	1	-1	25	0.9991	2.000	-2.901	37
500	1	-1	25	0.0009	2.000	-3.423	15
1000	1	-2	18	0.9991	2.000	-2.000	98
2500	1	-2	25	0.9991	2.000	-2.400	86
5000	1	-2	25	0.9975	2.000	-2.000	93
10000	3	-2	25	0.9991	2.000	-2.389	94
25000	5	-2	22	0.2953	2.000	-2.471	88

Table 12. Best hyperparameters of the semi-supervised experiment on MNIST (pca50) for the semi-supervised hybrid LM objective (SSL LM).

N_l	K	$\log_{10}\varepsilon$	R	λ	κ	$\log_{10}\gamma$	$\log_{10}\eta$	best epoch
100	3	-1	25	0.0009	0.0009	-0.544	-3.823	98
250	3	-1	25	0.9991	0.0009	-2.000	-2.858	29
500	4	-1	25	0.0009	0.0009	-0.841	-3.146	68
1000	5	-1	25	0.9991	0.9182	-2.000	-3.877	95
2500	5	-1	25	0.0398	0.0009	-2.000	-3.143	32
5000	5	-1	25	0.9991	0.0011	-2.000	-5.000	86
10000	5	-2	25	0.1227	0.0009	2.000	-2.274	64
25000	5	-2	25	0.7376	0.9991	2.000	-2.987	62

Table 13. Best hyperparameters of the semi-supervised experiment on MNIST (pca50) for the supervised hybrid CLL objective (SV CLL).

N_l	K	$\log_{10}\varepsilon$	R	λ	$\log_{10}\eta$	best epoch
100	1	-1	13	0.9991	-2.000	76
250	1	-1	20	0.0009	-3.640	93
500	1	-1	25	0.9982	-3.388	82
1000	1	-1	25	0.9991	-2.000	99
2500	3	-1	25	0.9988	-2.558	91
5000	5	-1	25	0.0009	-4.416	89
10000	5	-1	25	0.8049	-4.363	35
25000	5	-2	25	0.9991	-2.412	34

Table 14. Best hyperparameters of the semi-supervised experiment on MNIST (pca50) for the semi-supervised hybrid CLL objective (SSL CLL).

N_l	K	$\log_{10}\varepsilon$	R	λ	κ	$\log_{10}\eta$	best epoch
100	3	-1	25	0.0009	0.9836	-3.224	24
250	3	-1	23	0.0009	0.0817	-2.933	72
500	5	-1	25	0.9991	0.0009	-4.363	93
1000	5	-1	25	0.9989	0.5625	-5.000	99
2500	5	-1	25	0.0009	0.0009	-3.360	60
5000	5	-1	25	0.0330	0.0009	-4.000	52
10000	5	-1	25	0.0009	0.0057	-5.000	89
25000	5	-1	25	0.0009	0.9991	-3.395	75