Supplementary Material for an Efficient Optimization for Discriminative Latent Class Model

Armand Joulin*
INRIA
23, avenue d'Italie,
75214 Paris, France.
`armand.joulin@inria.fr`

Francis Bach*
INRIA
23, avenue d'Italie,
75214 Paris, France.
`francis.bach@inria.fr`

Jean Ponce*
Ecole Normale Supérieure
45, rue d'Ulm
75005 Paris, France.
`jean.ponce@ens.fr`

1 Approximation of the M-step

Using the relation between θ and ξ given by the M-step, we propose to divide our cost function into a term depending on α, another depending on (w, b) and a third one independent of θ. Taking the part of our cost function that depends on α, and replacing α by its expression, we get the function J_α:

$$J_\alpha(\xi) = \sum_{k=1}^{K}\sum_{m=1}^{M}\left(\frac{1}{N}\sum_{n\in A_m}\xi_{nk}\right)\log\left(\frac{1}{N}\sum_{n\in A_m}\xi_{nk}\right) - \sum_{k=1}^{K}\left(\frac{1}{N}\sum_{n=1}^{N}\xi_{nk}\right)\log\left(\frac{1}{N}\sum_{n=1}^{N}\xi_{nk}\right),$$

where A_m is the set of n such as $y_{nm} = 1$. Similarly with (w, b), we get the function J_{wb}:

$$J_{wb}(\xi) = \max_{w\in\mathbb{R}^{N\times K}, b\in\mathbb{R}^K} \frac{1}{N}\sum_{n=1}^{N}\xi_n(w^\top x_n + b) - \frac{1}{N}\sum_{n=1}^{N}\varphi(w^\top x_n + b) - \frac{\lambda}{2K}\|w\|_F^2,$$

where $\varphi(u) = \log(\sum_{k=1}^{K}\exp(u_k))$ is the log-sum exp function and ξ_n is the n-th row of ξ. Finally there is a third term independent of θ in $F(\xi, \theta)$:

$$J_C(\xi) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\xi_{nk}\log\xi_{nk}.$$

We call $F(\xi)$ the sum of $J_C(\xi)$, $J_{wb}(\xi)$ and $J_\alpha(\xi)$.

1.1 Expansion around symmetry

In this supplementary material, we show the second-order approximation of $J_C(\xi)$, $J_{wb}(\xi)$ and $J_\alpha(\xi)$ around the symmetric point where each observation has the same probability to be in each cluster, i.e. $p(z_n = k|x_n) = \frac{1}{K}$, $\xi_0 = \frac{1}{K}1_N 1_K^T$.

Second-order Taylor expansion of $J_C(\xi)$. Using the little-o notation defined as $a(x) = o(b(x))$ *if and only if* $\frac{a(x)}{b(x)} \to 0$ as $x \to 0$, we obtain:

$$J_C(\xi) = \log(K) - \frac{1}{2} - \frac{K}{2N}\text{tr}(\xi\xi^T) + O(\|\xi - \xi_0\|_F^3).$$

---

[*] WILLOW project-team, Laboratoire d'Informatique de l'Ecole Normale Supérieure, (ENS/INRIA/CNRS UMR 8548).

Indeed (we omit the $O(\|\xi-\xi_0\|_F^3)$ term):

$$\begin{aligned}
J_C(\xi) &= \log(K) + \frac{1}{N}(\log(K)-1)\sum_{n=1}^{N}\sum_{k=1}^{K}(\xi_{nk}-\frac{1}{K}) - \frac{K}{2N}\sum_{n=1}^{N}\sum_{k=1}^{K}(\xi_{nk}-\frac{1}{K})^2 \\
&= \log(K) - \frac{K}{2N}(\sum_{n=1}^{N}\sum_{k=1}^{K}\xi_{nk}^2 + \frac{NK}{K^2}) \\
J_C(\xi) &= \log(K) - \frac{1}{2} - \frac{K}{2N}\mathrm{tr}(\xi\xi^T)
\end{aligned}$$

Second-order Taylor expansion of $J_\alpha(\xi)$. Denoting by $Y \in \mathbb{R}^{N\times M}$, the matrix containing the y_{nm}, we obtain the expression:

$$\begin{aligned}
J_\alpha(\xi) = -\log(N) + \sum_{m=1}^{M}\frac{|A_m|}{N}\log(|A_m|) + \frac{K}{2N}\Big(\mathrm{tr}(\xi^T Y(Y^TY)^{-1}Y^T\xi) - \frac{1}{N}\mathrm{tr}(\xi 1_n 1_n^T\xi)\Big) \\
+O(\|\xi-\xi_0\|_F^3),
\end{aligned}$$

since $Y(Y^TY)^{-1}Y^T = \sum_{m=1}^{M}\frac{1}{|A_m|}1_{A_m}1_{A_m}^T$. Indeed:

$$\begin{aligned}
J_\alpha(\xi) &= \frac{1}{N}\sum_{m=1}^{M}|A_m|\log(\frac{|A_m|}{N}) - \frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N}\mathbb{1}_{n\in A_m}\log(\frac{|A_m|}{N})\sum_{k=1}^{K}(\xi_{nk}-\frac{1}{K}) \\
&+ \frac{K}{2N}\sum_{k=1}^{K}(\sum_{m=1}^{M}\sum_{n,l=1}^{N}\frac{1}{|A_m|}\mathbb{1}_{(n\in A_m)\cap(l\in A_m)}(\xi_{nk}-\frac{1}{K})(\xi_{lk}-\frac{1}{K}) - \frac{1}{N}(\sum_{n=1}^{N}\xi_{nk}-\frac{N}{K})^2) \\
&= -\log(N) + \sum_{m=1}^{M}\frac{|A_m|}{N}\log(|A_m|) + \frac{K}{2N}\sum_{k=1}^{K}(\sum_{m=1}^{M}\frac{1}{|A_m|}(\xi_k^T 1_{A_m})^2 - \frac{1}{N}(\sum_{n=1}^{N}\xi_{nk})^2) \\
&= -\log(N) + \sum_{m=1}^{M}\frac{|A_m|}{N}\log(|A_m|) + \frac{K}{2N}\Big(\sum_{m=1}^{M}\frac{1}{|A_m|}\mathrm{tr}(\xi^T 1_{A_m}1_{A_m}^T\xi) - \frac{1}{N}\mathrm{tr}(\xi 1_n 1_n^T\xi)\Big)
\end{aligned}$$

Second-order Taylor expansion of $J_{wb}(\xi)$. Solving a softmax regression problem is instable because the constant term b can go to the infinity. A common way to avoid this problem is to add a small regularization of b (in our case with a coefficient equals to 10^{-12}).

We note that $p(z_n = k|x_n) = \frac{1}{K}$ is equivalent to $w_k^T x_n + b_k = 0$. Thus we expand the log-sum-exp, φ around 0:

$$\varphi(x) = \log(K) + \frac{1}{K}x^T 1_K + \frac{1}{2K}\|x\|_F^2 - \frac{1}{2K^2}(x^T 1_K)^2 + O(\|x\|_F^3),$$

and substituting in $J_{wb}(\xi)$ yields:

$$\begin{aligned}
J_{wb}(\xi) = -\log(K) + \frac{K}{2N}\mathrm{tr}(\xi\Pi_K\xi^T) \\
- \frac{1}{2K}\min_{w,b}\Big[\frac{1}{N}\|(K\xi - Xw - b)\Pi_K\|_F^3 + \lambda\|w\|_F^2 + O(\|Xw+b\|_F^3)\Big],
\end{aligned}$$

where $\Pi_K = I - \frac{1}{K}1_K 1_K^T$ and $X = (x_1,\ldots,x_N)^\top$. Indeed:

$$
\begin{array}{rcl}
J_{wb}(\xi) & = & \displaystyle\max_{w\in\mathbb{R}^{P\times K},b\in\mathbb{R}^K} \frac{1}{N}\sum_{n=1}^{N}\xi_n(w^\top x_n+b) \\
& - & \displaystyle\frac{1}{N}\sum_{n=1}^{N}\Big[\log(K)+\frac{1}{K}(w^\top x_n+b)^T 1_K+\frac{1}{2K}\|w^\top x_n+b\|_F^2 \\
& - & \displaystyle\frac{1}{2K^2}((w^\top x_n+b)^T 1_K)^2\Big]-\frac{\lambda}{2K}\|w\|_F^2+O(\|Xw+b\|_F^3) \\
& = & \displaystyle -\log(K)+\max_{w,b}\frac{1}{N}\sum_{n=1}^{N}\Big[(\xi_n-\frac{1}{K}1_K^\top)(w^\top x_n+b) \\
& - & \displaystyle\frac{1}{2K}(\|w^\top x_n+b\|_F^2-((w^\top x_n+b)^T\frac{1}{K}1_K)^2) \\
& - & \displaystyle\frac{\lambda}{2K}\|w\|_F^2+O(\|Xw+b\|_F^3)\Big] \\
& = & \displaystyle -\log(K)\max_{w,b}\frac{1}{2KN}\sum_{n=1}^{N}\Big[2(K\xi_n)\Pi_K(w^\top x_n+b) \\
& - & \displaystyle \|\Pi_K(w^\top x_n+b)\|_F^2-\lambda\|w\|_F^2+O(\|Xw+b\|_F^3)\Big] \\
& = & \displaystyle -\log(K)+\frac{K}{2N}\mathrm{tr}(\xi\Pi_K\xi^T)-\frac{1}{2K}\min_{w,b}\Big[\frac{1}{N}\|K\xi\Pi_K \\
& - & \displaystyle (xw+b^T)\Pi_K\|_F^2+\lambda\|w\|_F^2+O(\|Xw+b\|_F^3)\Big]\Big].
\end{array}
$$

Due to the regularization in w and b, this cost function is Lipschitzian and the negligible terms in (w,b) becomes negligible terms in $\xi-\xi_0$. Therefore the minimization in respect to w and b corresponds to a multi-label classification problem with a square-loss [**?**, **?**, **?**]. This problem can be solve in a close form and leads to $b^*=\frac{1}{N}1_N1_N^T(K\xi-Xw))$ and to:

$$N\lambda w+x^T\Pi_N xw\Pi_K=Kx^T\Pi_N\xi,$$

and substituting in $J_{wb}(\xi)$ yields:

$$J_{wb}(\xi)=\log(K)-\frac{1}{2}+\frac{c(x)}{2N}+\frac{K}{2}\mathrm{tr}\Big[\xi\xi^T\big(\frac{1}{N}I-A(x,\lambda)\big)\Big]+O(\|\xi-\xi_0\|_F^3),$$

where $c(x)=\mathrm{tr}\big(1_N1_N^T(A(x,\lambda)-\Pi_N\big)$ and $A(x,\lambda)=\Pi_N(I-x(N\lambda I+x^T\Pi_N x)^{-1}x^T))\Pi_N$.

Second-order Taylor expansion. Finally combining these three terms and dropping the constant in ξ, we obtain:

$$F(\xi)=\frac{K}{2}\mathrm{tr}\Big[\xi\xi^T\Big(\frac{1}{N}\big(Y(Y^TY)^{-1}Y^T-\frac{1}{N}1_N1_N^T\big)-A(x,\lambda)\Big)\Big]. \qquad (1)$$

General case reformulation

We consider the problem:

$$\min\ \ \tfrac{1}{2}\,v^TQv \quad \text{s.t. } v\in\mathbb{R}^{NK},\ v\geq 0 \text{ and } (I_N\otimes 1_K^T)v=1_N. \qquad (2)$$

1.1.1 Problem reformulation

The set of completely positive matrices ($\mathcal{CP}_N$) is defined by:

$$\mathcal{CP}_N=\{M\in\mathbb{R}^{N\times N}|\exists p\in\mathbb{N}^*,\ \exists U\in\mathbb{R}^{N\times p},\ U\geq 0,\ M=UU^T\}$$

Denoting by $T = yy^T$, the $NK \times NK$ matrix, we consider its block matrix decomposition into $N \times N$ blocks of size $K \times K$:

$$T = \begin{bmatrix} T_{11} & T_{12} & \dots & T_{1N} \\ T_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ T_{N1} & \dots & \dots & T_{NN} \end{bmatrix}.$$

A $NK \times NK$ matrix T can be written as $T = yy^T$ *if and only if* it verifies the following conditions:

$$- \quad T \in \mathcal{CP}_{NK}, \tag{3}$$

$$- \quad \forall\, n, m \in \{1, \dots, N\},\ 1_K^T T_{nm} 1_K = 1, \tag{4}$$

$$- \quad \forall\, n, i, j \in \{1, \dots, N\},\ T_{ni} 1_K = T_{nj} 1_K, \tag{5}$$

$$- \quad \text{rank}(T) = 1. \tag{6}$$

Therefore the optimization problem (2) is equivalent to minimizing $\frac{1}{2}\text{tr}\ (TQ)$ over this set of constraints. However, these constraints do not define a convex set. In the next section we propose a convex relaxation based on the same idea as the simple case and a lowrank reformulation.

1.1.2 Relaxation

Dropping the rank condition leads to a matrix U such as $T = UU^T$ with $\forall (i, j),\ U_{ij} \geq 0$ and with at most a rank R (with $R > 1$). We note U_r the r-th column of U, U_r^n the n-th K-vector such as $U_r = (U_r^1, \dots, U_r^N)^T$ and $U^n = (U_1^n, \dots, U_R^n)$.

Since $T_{nm} = \sum_{r=1}^R U_r^n (U_r^m)^T$, conditions (4) and (5) can be replaced by conditions on U.

Condition (5) becomes for all m, $\sum_{r=1}^R 1^T U_r^m U_r^n = \sum_{r=1}^R 1^T U_r^n U_r^n$. Since, there are N such equalities for each U^n, this implies that for all m, $1^T U_r^m = 1^T U_r^n$. Adding $U \geq 0$, we have the new condition:

$$\forall (n, m) \leq N,\ \|U_r^m\|_1 U_r^n = \|U_r^n\|_1 U_r^n.$$

this condition means $\forall m \leq N,\ \|U_r^m\|_1 = \|U_r^n\|_1$, and therefore (4) can be reformulated:

$$\forall n \leq N,\ \sum_{r=1}^R (\|U_r^n\|_1)^2 = 1.$$

As in the simple case, we drop this condition by using a scale invariant cost function.
Finally, defining by $\mathcal{C}_1$, the set of constraints:

$$\mathcal{C}_1 = \{U_r \in \mathbb{R}^{NK} \mid U_r \geq 0,\ \forall n, m,\ \|U_r^n\|_1 = \|U_r^m\|_1\},$$

leads to a new formulation:

$$\min \ \ \tfrac{1}{2}\text{tr}(UD^{-1}U^TQ) \quad \text{s.t.} \quad U \in \mathbb{R}^{NK \times R} \quad \text{and} \quad \forall r,\ U_r \in \mathcal{C}_1. \tag{7}$$

where $D = \text{diag}((I_N \otimes 1_K)^T UU^T (I_N \otimes 1_K))$ and $\text{diag}(A)$ is the diagonal matrix with the diagonal of A.

1.2 Notes on the projection on $\mathcal{C}_1$

Remark on the projection. There is a linear function between the λ_n and a [**?**], whch yields that, for a given active set, $L(a)$ is a quadratic function of a. Since $L(a)$ is also continuous, $L(a)$ is piecewise quadratic. It means that for each segment there is we can evaluate the best a in close form. However, there are K^N different segments.
Complexity. For the binary search, the bottleneck of this projection is to sort the coefficients of each Z^n at the beginning. The overall complexity is therefore $O(NK\log(K))$.

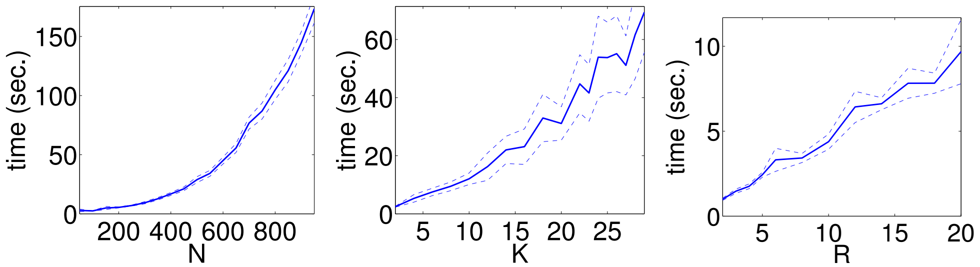

Figure 1: Running time as a function of N, K, and R.

2 Results

2.1 Running time

Empirically, we have verified this complexity on a toy example. Results are shown Fig. 1. We experiment the running time of our algorithm on 50 random matrices Q obtained with a uniform distribution over $[0, 1]$ for increasing values of N, K, and R.

2.2 Application to classification

Figure 2 shows all the results on the five binary classification tasks on *20 Newsgroups* dataset[1]. Since each document has 13312 entries, we set our degree of freedom at $df = 500$ and deduce from it the value of our regularization parameter λ. We use 50 random initializations for our algorithm. We compare our method with classifiers such as the linear SVM and the supervised Latent Dirichlet Allocation (sLDA) classifier of Blei et al. [**?**]. We also compare our results to those obtained by an SVM using on the features obtained with rank reducing methods such as the LDA of Blei et al. [**?**] and the PCA. For these models, we select their parameters with 5-fold cross-validation.

---

[1] http://people.csail.mit.edu/jrennie/

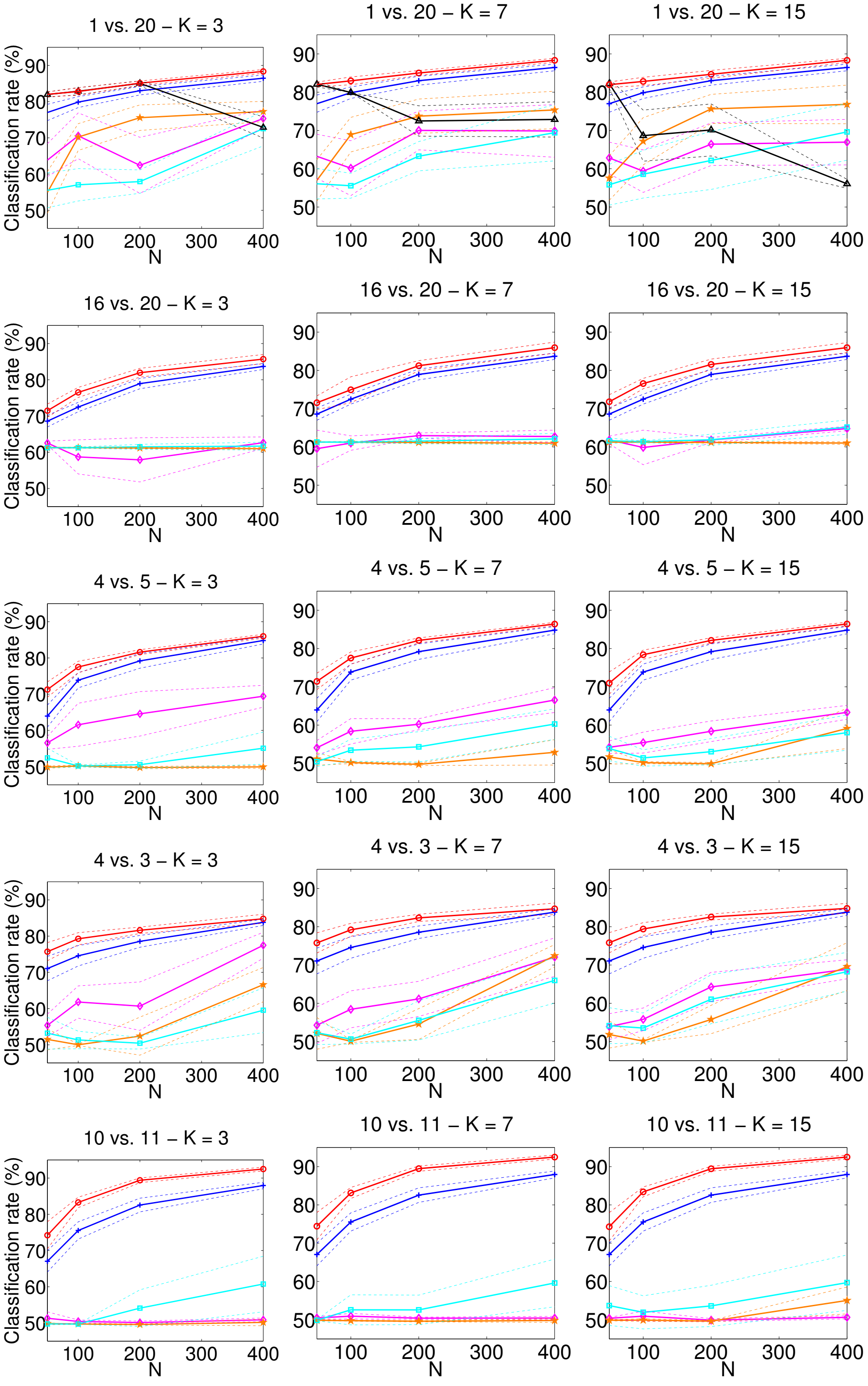

Figure 2: Classification rate for several binary classification tasks (from to bottom) and for different number of class K (or topics) (from left to right). (Same legend as in the paper).