Generalized Majorization-Minimization
Supplementary Material

Sobhan Naderi [1] **Kun He** [2] **Reza Aghajani** [3] **Stan Sclaroff** [4] **Pedro Felzenszwalb** [5]

In this supplementary material we will provide proofs for the two theorems that we presented in the main paper. We also provide more visualization of the models trained with G-MM and compare them with CCCP and EM, which we had to omit from the main paper due to space limitations.

1. Proof of Convergence

Proof of Theorem 1. First, we observe that the following inequality follows from the bound construction assumptions:

$$b_t(w_t) \leq b_t(w_{t-1}) \leq v_{t-1}, \tag{14}$$

where $v_t = b_t(w_t) - \eta d_t$. In particular, the first inequality holds because w_t minimizes b_t and the second inequality follows from (3). Summing (14) over $t = 1, ..., T$ and substituting the definition of v_t gives

$$\sum_{t=1}^{T} b_t(w_t) \leq \sum_{t=1}^{T} v_{t-1} = v_0 + \sum_{t=1}^{T-1} \big(b_t(w_t) - \eta d_t\big)$$

which implies

$$\eta \sum_{t=1}^{T} d_t \leq v_0 - b_T(w_T). \tag{15}$$

Recall that we set $v_0 = F(w_0)$, and let $F_* \in \mathbb{R}$ denote a finite global lower bound for F, and hence $b_T(w_T) \geq F_*$. The bound (15) then implies

$$\eta \sum_{t=1}^{\infty} d_t \leq F(w_0) - F_* < \infty,$$

which gives $\lim_{t\to\infty} d_t = 0$.

Next, recall that for every m-strongly convex function f, every x, y in the domain of f, and every subgradient $g \in \partial f(x)$, we have

$$f(y) \geq f(x) + g^T(y - x) + \frac{m}{2}||x - y||^2. \tag{16}$$

Substituting $f = b_t$, $x = w_t$, and $y = w_{t-1}$ in (16), and noting that the zero vector is a subgradient of b_t at w_t (because w_t is a minimizer of b_t), we obtain

$$\begin{aligned} ||w_t - w_{t-1}||^2 &\leq \frac{2}{m}\left(b_t(w_{t-1}) - b_t(w_t)\right) \\ &\leq \frac{2}{m}\left(b_{t-1}(w_{t-1}) - b_t(w_t)\right), \end{aligned} \tag{17}$$

where (3) is used in the second inequality. Summing (17) over $t = 2, .., T$, we obtain

$$\begin{aligned} \sum_{t=1}^{T} ||w_t - w_{t-1}||^2 &\leq b_1(w_1) - b_T(w_T) \\ &\leq F(w_1) - F_*, \end{aligned} \tag{18}$$

which implies

$$\lim_{t\to\infty} \|w_t - w_{t-1}\| = 0 \tag{19}$$

On the other hand, since $F(w_t) \leq b_t(w_t) \leq F(w_0)$ by (2), the sequence $\{w_t\}_t$ lies in the sublevel set $\{w \in \mathbb{R}^n | F(w) \leq F(w_0)\}$, which is assumed to be a compact set. To show that a sequence that is contained in a compact set converges, one needs to prove that all its converging subsequences have the same limit. For $\{w_t\}_t$, this follows from (19), and therefore $\{w_t\}_t$ converges to a limit $w^\dagger$. □

Proof of Theorem 2. We prove this theorem by contradiction. Suppose $\nabla F(w^\dagger) \neq 0$. This implies that there exists a unit vector $u \in \mathbb{R}^d$ such that the directional derivative of F along u is positive at $w^\dagger$, *i.e.* $\nabla F(w^\dagger) \cdot u > 2c$ for some $c > 0$. Since F is continuously differentiable, $\nabla F \cdot u$ is continuous at $w^\dagger$, and hence

$$\nabla F(w) \cdot u > c, \qquad \forall w \in B_{2\delta}(w^\dagger), \tag{20}$$

for all small enough $\delta > 0$, where $B_r(x)$ denotes an open ball around x with radius r. We fix a $\delta > 0$ that satisfies (20), as well as the bound

$$\delta < \frac{2c}{M}. \tag{21}$$

* The paper was written when Kun He was at Boston University. [1]Google Research [2]Facebook Reality Labs [3]University of California San Diego [4]Boston University [5]Brown University. Correspondence to: Sobhan Naderi <sobhan@google.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

We also fix an $\epsilon > 0$ that satisfies

$$\epsilon < c\delta - \frac{M}{2}\delta^2, \tag{22}$$

which is possible because of (21). The reason for this will be clear shortly.

Now recall by Theorem 1 that $w_t \to w^\dagger$ and $d_t \to 0$, as $t \to \infty$, so we can pick $T > 0$ large enough such that

$$|w_T - w^\dagger| < \delta \tag{23}$$

and

$$d_T = b(w_T) - F(w_t) < \epsilon \tag{24}$$

Now define the function g to be the restriction of F on a line parallel to u that passes through w_T (see Figure 1), that is

$$g(z) = F(w_T + zu), \qquad z \in \mathbb{R}.$$

It is easy to see that g is continuously differentiable with

$$g'(z) = \nabla F(w_T + zu) \cdot u.$$

In particular, the bound (20) implies

$$g'(z) > c, \qquad z \in (0, \delta). \tag{25}$$

This is because for every $z \in (0, \delta)$,

$$w_T + zu \in B_\delta(w_T) \subset B_{2\delta}(w^\dagger).$$

An application of Taylor Expansion Theorem of order $n = 0$ on g around $z = 0$ shows that there exits a $z_* \in (0, \delta)$ such that

$$g(\delta) = g(0) + g'(z_*)\delta > g(0) + c\delta,$$

where we used $g'(z_*) > c$ by (25). Substituting definitions of $g(0)$ and $g'(0)$ in the display above, we obtain the bound

$$F(w_*) > F(w_T) + c\delta, \quad w_* = w_T + \delta u. \tag{26}$$

On the other hand, since b_T is a smooth function with its minimum at w_T and its Hessian $\nabla^2 b_T$ bounded by MI, second order Taylor expansion of b_T around w_T gives

$$b_T(w) \leq b_T(w_T) + \nabla b_T(w_T) \cdot (w - w_T) + \frac{M}{2}\|w - w_T\|^2,$$

and in particular, for $w = w_* = w_T + \delta u$,

$$b_T(w_*) \leq b_T(w_T) + \frac{M}{2}\delta^2. \tag{27}$$

Combining the bounds (24)-(27) and the choice (22) of ϵ, we have

$$\begin{aligned} b_T(w_*) - F(w_*) &\leq [b_T(w_T) - F(w_T)] + \frac{M}{2}\delta^2 - c\delta \\ &\leq \epsilon + \frac{M}{2}\delta^2 - c\delta \\ &< 0, \end{aligned}$$

which contradicts the fact that b_T is an upper bound for F. This completes the proof. $\square$

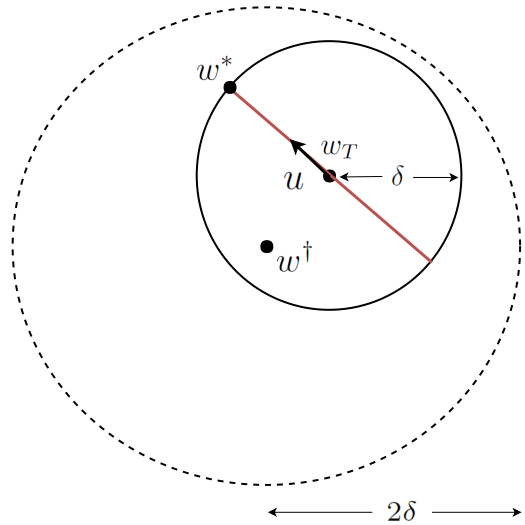


Figure 1. An illustration of quantities defined in the proof of Theorem 2

2. k-means Clustering

Figure 2 visualizes the result of k-means and G-MM (with random bounds) on the D-31 dataset (Veenman et al., 2002), from the same initialization. G-MM finds a near perfect solution, while in standard k-means, many clusters get merged incorrectly or die off. Dead clusters are those which do not get any points assigned to them. The update rule (M-step of k-means algorithm) collapses the dead clusters on to the origin.

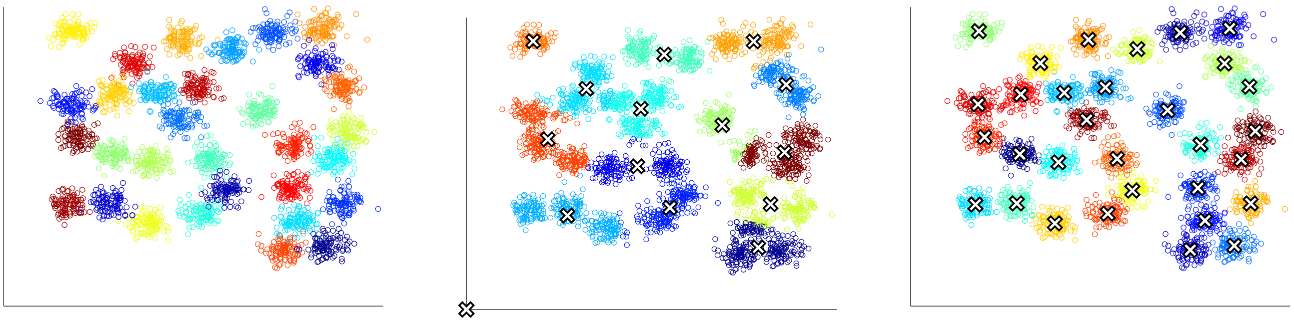
3. LS-SVM for Mammal Image Classification

We provide additional experimental results on the mammals dataset. Figure 3 shows example training images and the final imputed latent object locations by three algorithms: CCCP (red), G-MM random (blue), and G-MM biased (green). The initialization is *top-left*.

In most cases CCCP fails to update the latent locations given by initialization. The two G-MM variants, however, are able to update them significantly and often localize the objects in training images correctly. This is achieved *only* with image-level object category annotations, and with a very bad (even adversarial) initialization.

References

Veenman, C. J., Reinders, M., and Backer, E. A maximum variance cluster algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002.



(a) ground truth (b) k-means (c) generalized-MM

Figure 2. Visualization of clustering solutions on the D31 dataset (Veenman et al., 2002) from identical initializations. Random partition initialization scheme is used. (a) color-coded ground-truth clusters. (b) solution of k-means. (c) solution of G-MM. The white crosses indicate location of the cluster centers. Color codes match up to a permutation.

Figure 3. Example training images from the mammals dataset, shown with final imputed latent object locations by three algorithms: CCCP (red), G-MM random (blue), G-MM biased (green). Initialization: *top-left*.