Appendix: Sharing Features in Multi-class Boosting via Group Sparsity

Sakrapee Paisitkriangkrai, Chunhua Shen, Anton van den Hengel
The University of Adelaide, Australia

March, 2012

In this document we provide a complete derivation for multi-class boosting with group sparsity and a full explanation of ADMM algorithm presented in the main paper.

1 Multi-class boosting with group sparsity

We first provide the derivation for multi-class logistic loss with $\ell_{1,2}$-norm. We then show the difference between our boosting with $\ell_{1,2}$-norm and $\ell_{1,\infty}$-norm. We then briefly discuss our group sparsity-based boosting for any general convex loss.

1.1 Multi-class logistic loss

As discussed, the learning problem for logistic loss in an $\ell_{1,2}$ regularization framework can be expressed as,

$$\begin{aligned} \min_{W,V,\boldsymbol{\rho}} \quad & \frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k}\log\big(1+\exp\left(-\rho_{ir}\right)\big)+\nu\|V\|_{1,2} \\ \text{s.t.} \quad & \rho_{ir}=H_{i:}\boldsymbol{w}_{y_i}-H_{i:}\boldsymbol{w}_r,\forall i,\forall r, \text{ and } V=W; W\geq 0. \end{aligned} \tag{1}$$

Here we introduce the auxiliary variables, $\boldsymbol{\rho}$, and additional constraints, $V=W$, to obtain the meaningful dual formulation. The Lagrangian of (1) can be written as

$$L=\frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k}\log\big(1+\exp\left(-\rho_{ir}\right)\big)+\nu\|V\|_{1,2}-\sum_{i,r}U_{ir}(\rho_{ir}-H_{i:}\boldsymbol{w}_{y_i}+H_{i:}\boldsymbol{w}_r)-\langle Q,\nu W-\nu V\rangle-\langle P,W\rangle\,, \tag{2}$$

with $U\geq 0$ and $P\geq 0$. At optimum the first derivative of the Lagrangian w.r.t. each row of W must be zeros

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{w}_r}=\mathbf{0}\rightarrow & \sum_{i,r=y_i}\left(\textstyle\sum_l U_{il}\right)H_{i:}-\textstyle\sum_i U_{ir}H_{i:}=P_{:r}-\nu Q_{:r} \\ & \rightarrow \textstyle\sum_i\left[\delta_{r,y_i}\left(\sum_l U_{il}\right)-U_{ir}\right]H_{i:}\geq -\nu Q_{:r} \end{aligned} \tag{3}$$

for $\forall r$. Take infimum over the primal variable, ρ_{ir},

$$\frac{\partial L}{\partial \rho_{ir}}=0\rightarrow \rho_{ir}=-\log\left(\frac{-mkU_{ir}}{mkU_{ir}-1}\right),\forall i,\forall r. \tag{4}$$

and

$$\inf_{\rho_{ir}} L=\frac{1}{mk}\sum_{ir}-(1+mkU_{ir})\log\left(1+mkU_{ir}\right)-mkU_{ir}\log\left(-mkU_{ir}\right) \tag{5}$$

By reversing the sign of U and using the fact that the convex conjugate of $\|V_{j:}\|_2$ is the indicator function of the dual norm unit ball [2]. the Lagrange dual can be written as,

$$\begin{aligned} \max_{U,Q} \quad & -\frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k}\Big[mkU_{ir}\log\left(mkU_{ir}\right)+\left(1-mkU_{ir}\right)\log\left(1-mkU_{ir}\right)\Big] \\ \text{s.t.} \quad & \textstyle\sum_i\big[\delta_{r,y_i}\left(\sum_l U_{il}\right)-U_{ir}\big]H_{i:} \leq \nu Q_{:r}, \forall r; \text{ and } \|Q_{j:}\|_2 \leq 1, \forall j. \end{aligned} \tag{6}$$

1.2 Multi-class boosting with $\ell_{1,\infty}$-norm

The derivation here is very similar to what we derived in our main paper (multi-class boosting with $\ell_{1,2}$-norm). In this section, we give an example of multi-class hinge loss. However, it would be very straightforward to apply this to multi-class logistic loss. For the sake of clarity, we rewrite both $\ell_{1,2}$ and $\ell_{1,\infty}$ objective functions below.

$$\begin{aligned} \min_{W,V,\boldsymbol{\xi}} \quad & \sum_{i=1}^{m}\xi_i+\nu\|V\|_{1,2} \\ \text{s.t.} \quad & \delta_{r,y_i}+H_{i:}\boldsymbol{w}_{y_i} \geq 1+H_{i:}\boldsymbol{w}_r-\xi_i, \forall i,r; \text{ and } V=W; W\geq 0; \boldsymbol{\xi}\geq 0. \end{aligned} \tag{7}$$

$$\begin{aligned} \min_{W,V,\boldsymbol{\xi}} \quad & \sum_{i=1}^{m}\xi_i+\nu\|V\|_{1,\infty} \\ \text{s.t.} \quad & \delta_{r,y_i}+H_{i:}\boldsymbol{w}_{y_i} \geq 1+H_{i:}\boldsymbol{w}_r-\xi_i, \forall i,r; \text{ and } V=W; W\geq 0; \boldsymbol{\xi}\geq 0. \end{aligned} \tag{8}$$

The only difference between (7) and (8) is in the regularisation term. We can derive the Lagrange dual similar to the case of $\ell_{1,2}$-norm. The Lagrangian of (8) can be written as,

$$L=\sum_{i=1}^{m}\xi_i+\nu\|V\|_{1,\infty}-\sum_{i,r}U_{ir}(\delta_{r,y_i}+H_{i:}\boldsymbol{w}_{y_i}-1-H_{i:}\boldsymbol{w}_r+\xi_i)-\langle Q,\nu W-\nu V\rangle-\langle P,W\rangle,$$

with $U\geq 0$ and $P\geq 0$. For $\ell_{1,\infty}$-norm, the infimum over the primal variables V can be expressed as,

$$\begin{aligned} \inf_V L &= \inf_V -\nu\langle Q,V\rangle+\nu\|V\|_{1,\infty} \\ &= -\nu\textstyle\sum_j \sup_{V_{j:}}\langle Q_{j:},V_{j:}\rangle+\nu\sum_j\|V_{j:}\|_\infty \\ &= -\nu\textstyle\sum_j\Big[\sup_{V_{j:}} Q_{j:}^\top V_{j:}-\|V_{j:}\|_\infty\Big] \\ &= -\nu\textstyle\sum_j\begin{cases}0 & \text{if } \|Q_{j:}\|_1\leq 1,\forall j,\\ \infty & \text{otherwise.}\end{cases} \end{aligned} \tag{9}$$

Here we make use of the fact that,

$$f^*(y)=\sup_x(y^\top x-\|x\|)=\begin{cases}0 & \text{if } \|y\|_*\leq 1,\\ \infty & \text{otherwise,}\end{cases} \tag{10}$$

where $\|\cdot\|_*$ is dual norm of $\|\cdot\|$ [1]. Hence we can derive its corresponding dual as,

$$\begin{aligned} \min_{U,Q} \quad & \sum_{i,r}U_{ir}\delta_{r,y_i} \\ \text{s.t.} \quad & \textstyle\sum_i(\delta_{r,y_i}-U_{ir})H_{ij}\leq\nu Q_{:r}, \forall r; \text{ and } \sum_r U_{ir}=1, \forall i; U\geq 0; \text{ and } \|Q_{j:}\|_1\leq 1, \forall j. \end{aligned} \tag{11}$$

[1] We note here that ℓ_p norm in primal corresponds to ℓ_q norm in dual with $1/p+1/q=1$. For example, the Euclidean norm, $\|\cdot\|_2$ is dual to itself and the ℓ_1-norm, $\|\cdot\|_1$ is dual to the ℓ_∞-norm, $\|\cdot\|_\infty$.

Implementation The only difference between the dual of $\ell_{1,2}$-regularized and $\ell_{1,\infty}$-regularized boosting is in the constraint part, *i.e.*, $\|Q_{j:}\|_2 \leq 1, \forall j$ for $\ell_{1,2}$-norm and $\|Q_{j:}\|_1 \leq 1, \forall j$ for $\ell_{1,\infty}$-norm. From the given MultiBoost$^{\text{group}}$algorithm in the main paper, two modifications need to be made: 1) replace ℓ_2-norm in the stopping criterion with ℓ_1-norm; 2) solve (8) instead of (7) in the step 4 in the algorithm.

1.3 A more general convex loss

In this section, we generalize our idea to any convex loss functions. We define $l(\cdot)$ as a smooth convex function. We define the margin as the pairwise difference of prediction scores. The general $\ell_{1,2}$-regularized optimization problem we want to solve is,

$$\begin{aligned} \min_{W,V,\boldsymbol{\rho}} \quad & \sum_{i=1}^{m}\sum_{r=1}^{k} l(\rho_{ir}) + \nu\|V\|_{1,2} \\ \text{s.t.} \quad & \rho_{ir} = H_{i:}\boldsymbol{w}_{y_i} - H_{i:}\boldsymbol{w}_r, \forall i, \forall r, \text{ and } V = W; W \geq 0. \end{aligned} \tag{12}$$

The Lagrangian of (12) can be written as,

$$L = \sum_{i=1}^{m}\sum_{r=1}^{k} l\left(\rho_{ir}\right) + \nu\|V\|_{1,2} \ - \sum_{i,r} U_{ir}(\rho_{ir} - H_{i:}\boldsymbol{w}_{y_i} + H_{i:}\boldsymbol{w}_r) - \langle Q, \nu W - \nu V\rangle - \langle P, W\rangle, \tag{13}$$

Following our derivation for multi-class logistic loss, the Lagrange dual can be written as,

$$\begin{aligned} \min_{U,Q} \quad & \sum_{i=1}^{m}\sum_{r=1}^{k} l^*(-U_{ir}) \\ \text{s.t.} \quad & \textstyle\sum_i \left[\delta_{r,y_i}\left(\sum_l U_{il}\right) - U_{ir}\right] H_{i:} \leq \nu Q_{:r}, \forall r; \text{ and } \|Q_{j:}\|_2 \leq 1, \forall j. \end{aligned} \tag{14}$$

where $l^*(\cdot)$ is the Fenchel dual function of $l(\cdot)$. Through the KKT condition, the relationship between the dual variable U and the primal variable ρ,

$$U_{ir} = -l'(\rho_{ir}), \tag{15}$$

holds at optimality. It is important to note here the difference between MultiBoost$^{\ell_1}$[3] and our works. Although our dual variables, U, have the same expression as theirs, *i.e.*, each dual variable is defined as the negative gradient of the loss at ρ_{ir}, the solution of our primal variables, W, are different from theirs. Unlike our work, MultiBoost$^{\ell_1}$does not enforce group sparsity. Hence, their algorithm fails to exploit the existence of structural features.

2 ADMM Implementation

2.1 Parallel optimization for FAST boosting

In the main paper, we design a boosting algorithm for optimizing the logistic loss function:

$$\frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k} \log\left(1 + \exp\left(H_{i:}\boldsymbol{w}_r - H_{i:}\boldsymbol{w}_{y_i}\right)\right). \tag{16}$$

We regularize the above logistic loss with a mixed-norm $\ell_{1,2}$ regularization. The learning problem is expressed as,

$$\begin{aligned} \min_{W,V,\boldsymbol{\rho}} \quad & \frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k} \log\left(1 + \exp\left(-\rho_{ir}\right)\right) + \nu\|V\|_{1,2} \\ \text{s.t.} \quad & \rho_{ir} = H_{i:}\boldsymbol{w}_{y_i} - H_{i:}\boldsymbol{w}_r, \forall i, \forall r \text{ and } V = W; W \geq 0. \end{aligned} \tag{17}$$

Since $\ell_{1,2}$-norm is not differentiable everywhere. To solve (17), we apply Alternating Direction Method of Multipliers (ADMM) [1]. ADMM formulates the original problem as the following,

$$\begin{aligned} \min_{W,Z} \quad & f(W) + g(Z) \\ \text{s.t.} \quad & W = Z. \end{aligned} \tag{18}$$

Here $f(W)$ is the logistic loss (16) and $g(Z)$ is $\nu\|W\|_{1,2}$. As in the method of multipliers, we form the augmented Lagrangian,

$$L_\lambda = f(W) + g(Z) + \langle U, W - Z\rangle + \frac{\lambda}{2}\|W - Z\|_2. \tag{19}$$

Here λ is the augmented Lagrangian parameter ($\lambda > 0$). The method of multipliers for (18) has the form,

$$(W^{s+1}, Z^{s+1}) = \operatorname*{argmin}_{W,Z} L_\lambda(W, Z, U^s) \tag{20}$$

$$U^{s+1} = U^s + \lambda(W^{s+1} - Z^{s+1}). \tag{21}$$

Here the Lagrangian is minimized jointly with respect to both W and Z variables. Since it is expensive to solve a joint minimization in (20), both primal variables (W and Z) are updated in an alternating fashion. This alternate update scheme is known as ADMM. ADMM consists of the following iterations,

$$W^{s+1} = \operatorname*{argmin}_{W} L_\lambda(W, Z^s, U^s) \tag{22}$$

$$Z^{s+1} = \operatorname*{argmin}_{Z} L_\lambda(W^{s+1}, Z, U^s) \tag{23}$$

$$U^{s+1} = U^s + \lambda(W^{s+1} - Z^{s+1}). \tag{24}$$

We can rewrite (22) and (23) as,

$$W^{s+1} = \operatorname*{argmin}_{W} \frac{1}{mk}\sum_{i=1}^{m}\sum_{r=1}^{k}\log\left(1 + \exp\left(-\rho_{ir}\right)\right) + (U^s)^\top W + \frac{\lambda}{2}\|W - Z^s\|_2^2 \tag{25}$$

$$Z^{s+1} = \operatorname*{argmin}_{Z} \nu\|Z\|_{1,2} - (U^s)^\top Z + \frac{\lambda}{2}\|W^{s+1} - Z\|_2^2. \tag{26}$$

where $\rho_{ir} = H_{i:}\boldsymbol{w}_r - H_{i:}\boldsymbol{w}_{y_i}$. Since (25) is now smooth and differentiable everywhere, a quasi-Newton method such as L-BFGS-B can be used to efficiently solve (25). For (26), a closed-form solution exists and it can be computed through subdifferential calculus [1]. The solution is known as a block soft thresholding,

$$Z_{j:}^{s+1} = \mathcal{S}_{\lambda/\rho}(W_{j:}^{s+1} + U_{j:}^s), \forall j, \tag{27}$$

where $\mathcal{S}$ is a vector soft thresholding operator defined as $\mathcal{S}_\kappa(a) = (1 - \kappa/\|a\|_2)_+ a$.

2.2 Distributed optimization

We describe here how to exploit distributed computing in ADMM to speed up the training time of our MultiBoost$^{\text{group}}$ and MultiBoost$^{\text{group}}_{\text{FAST}}$. Note that this approach is also applicable to MultiBoost$^{\ell_1}$ proposed in [3]. In order to solve the problem in a distributed fashion, we first separate the loss function across Q blocks of data. We redefine our problem as,

$$\begin{aligned} \min_{W,Z} \quad & \sum_{q=1}^{Q} l_q(W_q) + \nu\|Z\|_{1,2} \\ \text{s.t.} \quad & W_q - Z = 0, q = 1, \cdots, Q, \end{aligned} \tag{28}$$

where l_q refers to the loss function for the q-th block of data. Similar to the previous section, ADMM considers the following iterations,

$$W_q^{s+1} = \underset{W_q}{\operatorname{argmin}} L_\lambda(W_q, Z^s, U^s), \forall q; \tag{29}$$

$$Z^{s+1} = \underset{Z}{\operatorname{argmin}} L_\lambda(W_1^{s+1}, \cdots, W_Q^{s+1}, Z, U_1^s, \cdots, U_Q^s); \tag{30}$$

$$U_q^{s+1} = U_q^s + \lambda(W_q^{s+1} - Z^{s+1}), \forall q, \tag{31}$$

where λ is the augmented Lagrangian parameter ($\lambda > 0$). The resulting ADMM algorithm for (29) and (30) is

$$W_q^{s+1} = \underset{W_q}{\operatorname{argmin}}\ l_q(W_q) + (U_q^s)^\top W_q + \frac{\lambda}{2}\|W_q - Z^s\|_2^2, \forall q, \tag{32}$$

$$Z^{s+1} = \underset{Z}{\operatorname{argmin}}\ \nu\|Z\|_{1,2} + \sum_{q=1}^{Q}\Big(-\left(U_q^s\right)^\top Z + \frac{\lambda}{2}\|W_q^{s+1} - Z\|_2^2\Big), \tag{33}$$

$$= \mathcal{S}_{\lambda/\rho Q}(\bar{W}_{j:}^{s+1} + \bar{U}_{j:}^s), \forall j; \tag{34}$$

$$U_q^{s+1} = U_q^s + \lambda(W_q^{s+1} - Z^{s+1}), \forall q, \tag{35}$$

where $\bar{W}^{s+1} = \frac{1}{Q}\sum_{q=1}^{Q} W_q^{s+1}$ and $\bar{U}^s = \frac{1}{Q}\sum_{q=1}^{Q} U_q^s$. Here we assume that $\sum_{q=1}^{Q} m_q = m$, *i.e.*, the sum of the number of samples in each block is equal to the total number of samples. The first step, (32), can be carried out independently in parallel for each block of data. In other words, we distribute (32) to each thread or processor. The second step, (34), gathers variables computed in (32) to form the average. After the final step, (35), the value of U_q^{s+1} is then distributed to the subsystems.

For both hinge loss and logistic regression, we can rewrite (32) as,

$$W_q^{s+1} = \underset{W_q}{\operatorname{argmin}} \frac{1}{m_q}\sum_{i=1}^{m_q}\xi_i + (U_q^s)^\top W_q + \frac{\lambda}{2}\|W_q - Z^s\|_2^2; \tag{36}$$

$$W_q^{s+1} = \underset{W_q}{\operatorname{argmin}} \frac{1}{m_q k}\sum_{i=1}^{m_q}\sum_{r=1}^{k}\log\Big(1 + \exp\left(-\rho_{ir}\right)\Big) + (U_q^s)^\top W_q + \frac{\lambda}{2}\|W_q - Z^s\|_2^2; \tag{37}$$

References

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations & Trends in Mach. Learn.*, 3(1), 2011.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] C. Shen and Z. Hao. A direct formulation for totally-corrective multi-class boosting. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2011.