Alternating Direction Methods with Bregman Divergences

Huahua Wang **Arindam Banerjee**

Department of Computer Science & Engineering
University of Minnesota, Twin Cities, MN
{huwang,banerjee}@cs.umn.edu

1 Introduction

In recent years, minimizing composite objective functions have been widely studied in machine learning [11, 6, 5]. Formally, composite objective minimization problem has the following form:

$$\min_{\mathbf{w}\in\Omega} h(\mathbf{w}) \triangleq f(\mathbf{w}) + g(\mathbf{w}) \,, \tag{1}$$

where f, g are convex functions and Ω is a convex set. Provided it is easy to compute the (sub)gradients of both functions, a simple method to solve (1) is (sub)gradient descent,

$$\mathbf{w}_{t+1} = \Pi_{\mathbf{w}\in\Omega}\big[\mathbf{w}_t - \frac{1}{\eta} h'(\mathbf{w}_t)\big] \,, \tag{2}$$

where $\eta > 0$ is the step size, $h'(\mathbf{w}_t)$ is the (sub)gradient at $\mathbf{w}_t$, and $\Pi_{\mathbf{w}\in\Omega}(\mathbf{v}) = \operatorname{argmin}_{\mathbf{w}\in\Omega} \frac{1}{2}\|\mathbf{w}-\mathbf{v}\|_2^2$ denotes the projection onto the feasible set Ω in terms of the Euclidean distance. Alternatively, gradient descent can be reformulated in the form of a proximal gradient method [4]:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\Omega} \langle h'(\mathbf{w}_t), \mathbf{w}\rangle + \frac{\eta}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 \,, \tag{3}$$

where $\eta > 0$ is the step size and $h'(\mathbf{w}_t)$ is the (sub)gradient at $\mathbf{w}_t$. If there exists efficient algorithm for the projection, gradient descent is an efficient and has been successfully applied in large scale optimization. However, the use of Euclidean distance as a proximal function may be unsuitable for certain functions and constraint sets, yielding inefficient updates. For example, for loss functions based on entropy with constraint set being the unit simplex, a KL-divergence based proximal function is more appropriate. To accommodate such underlying structures in the problem, general proximal functions based on Bregman divergences are used in mirror descent algorithms (MDA) [1], where the updates are given by:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\Omega} \langle h'(\mathbf{w}_t), \mathbf{w}\rangle + \eta B_\phi(\mathbf{w}, \mathbf{w}_t) \,, \tag{4}$$

where B_ϕ is a Bregman divergence [1]. In particular, if B_ϕ is the KL-divergence, MDA leads to exponentiated gradient or multiplicative update algorithms [10] in contrast to additive update in gradient descent. Mirror descent can be considered a variant of the proximal method with Bregman divergence or D-function (PMD) [3]:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in\Omega} h(\mathbf{w}) + \eta B_\phi(\mathbf{w}, \mathbf{w}_t) \,. \tag{5}$$

For composite objective with a differentiable function and a simple nonsmooth function, composite objective mirror descent (COMID) [5] only linearizes the differentiable function $f$and includes forward-backward splitting (FOBOS) [6] as a special case where the Bregman divergence is a quadratic function. For simple enough constraints, MDA may be able to yield efficient projections by carefully choosing the Bregman divergence. In general, however, the full projection requires an inner loop algorithm, leading to a double loop algorithm for solving (1) [12].

In this paper, we consider the composite objective optimization subject to an equality constraint such that

$$\min_{\mathbf{x}\in\mathcal{X},\mathbf{z}\in\mathcal{Z}} f(\mathbf{x}) + g(\mathbf{z}) \quad \text{s.t.} \quad A\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c} \,, \tag{6}$$

where $A \in \mathbb{R}^{m\times n_1}, \mathbf{B} \in \mathbb{R}^{m\times n_2}, \mathbf{c} \in \mathbb{R}^m$, $\mathbf{x} \in \mathcal{X} \in \mathbb{R}^{n_1\times 1}, \mathbf{z} \in \mathcal{Z} \in \mathbb{R}^{n_2\times 1}$, $\mathcal{X}$ and $\mathcal{Z}$ are convex sets. Compared to (1), the equality constraint introduces splitting variables and thus splits the functions and the constraints set Ω into simpler constraints $\mathcal{X}, \mathcal{Z}$. Our algorithms basically solve the following two subproblems: (1) $\mathbf{x}$-update involving $f(\mathbf{x})$ and $\mathcal{X}$; (2) $\mathbf{z}$-update involving $g(\mathbf{x})$ and $\mathcal{Z}$. If the two subproblems can be solved efficiently, we can avoid the full projection in MDA and COMID which involves both the composite objective and the intersection of $\mathcal{X}$ and $\mathcal{Z}$, i.e., Ω. The divide-and-conquer strategy is particularly useful for composite objective with different structures, e.g., entropy loss function plus nonsmooth function, and constraints set which is an intersection of simple constraints, e.g., linear constraints and doubly stochastic matrix constraints. The full projection onto linear constraints in MDA requires solving a linear program in each iteration. However, if introducing a slack variable to separate linear inequality from linear equality, the projection onto either of them is trivial [2]. For doubly stochastic matrix, the full projection requires alternating projections like Sinkhorn algorithm [13], but the projection-free updates can be obtained using splitting variables which will show in Section 3.

(6) can be solved by the well known alternating direction method (ADM) [2], which has been shown to have a $O(1/T)$ convergence rate [14, 8, 7]. In ADM, both $\mathbf{x}$ and $\mathbf{z}$ updates amount to solving proximal minimization problems using the quadratic penalty term, thereby preventing the full utilization of the structures underlying the function and constraints set. In this paper, we propose Bregman ADMs where Bregman divergences can be used as proximal functions in ADM updates. Thus, Bregman ADMs generalize ADMs similar to how proximal methods generalize gradient descent. In Bregman ADMs, $\mathbf{x}$ and $\mathbf{z}$ updates can take the form of MDA (4) or PMD (5). In particular, Bregman ADM updates become alternating additive updates when using quadratic penalty and alternating multiplicative updates when using KL divergence. As an illustrative example, we consider minimization problems over doubly stochastic matrices and show that Bregman ADMs lead to an efficient single-loop algorithm for such problems.

2 Alternating Direction Methods with Bregman Divergences

In each iteration, ADM consists of the following three updates:

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \langle \mathbf{y}_t, A\mathbf{x} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\rangle + \frac{\rho}{2}\|A\mathbf{x} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|^2 , \tag{7}$$

$$\mathbf{z}_{t+1} = \operatorname{argmin}_{\mathbf{z}\in\mathcal{Z}} g(\mathbf{z}) + \langle \mathbf{y}_t, A\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z} - \mathbf{c}\rangle + \frac{\rho}{2}\|A\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2 , \tag{8}$$

$$\mathbf{y}_{t+1} = \mathbf{y}_t + \tau\rho(A\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}) . \tag{9}$$

where $\mathbf{y}$ is dual variable and $\rho > 0$ is penalty parameter. The $\mathbf{x}$ and $\mathbf{z}$ updates resemble proximal minimization problem using quadratic term, which limits ADM to exploit the underlying structures. Similar to MDA, Bregman ADM will replace the quadratic terms in ADM by a Bregman divergence term. More specifically, one version of Bregman ADM simply linearizes the objective and adds a Bregman divergence in (7) and/or (8), leading to the following updates:

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \langle F(\mathbf{x}_t), \mathbf{x}\rangle + \eta_{\mathbf{x}} B_{\phi_{\mathbf{x}}}(\mathbf{x}, \mathbf{x}_t) , \tag{10}$$

$$\mathbf{z}_{t+1} = \operatorname{argmin}_{\mathbf{z}\in\mathcal{Z}} \langle G(\mathbf{z}_t), \mathbf{z}\rangle + \eta_{\mathbf{z}} B_{\phi_{\mathbf{z}}}(\mathbf{z}, \mathbf{z}_t) , \tag{11}$$

where $F(\mathbf{x}_t), G(\mathbf{z}_t)$ are linearizations of objectives in (7) and (8), i.e.,

$$F(\mathbf{x}_t) = f'(\mathbf{x}_t) + A^T\{\mathbf{y}_t + \rho(A\mathbf{x}_t + \mathbf{B}\mathbf{z}_t - \mathbf{c})\} ,$$
$$G(\mathbf{z}_t) = g'(\mathbf{z}_t) + \mathbf{B}^T\{\mathbf{y}_t + \rho(A\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c})\} ,$$

and $f'(\mathbf{x}_t) \in \partial f(\mathbf{x}_t), g'(\mathbf{z}_t) \in \partial g(\mathbf{z}_t)$. (10) and (11) are MDA updates, making Bregman ADM behave like alternating mirror descent. If both Bregman divergences are quadratic function, Bregman ADM leads to the following alternating additive updates:

$$\mathbf{x}_{t+1} = \Pi_{\mathbf{x}\in\mathcal{X}}[\mathbf{x}_t - \frac{1}{\eta_{\mathbf{x}}} F(\mathbf{x}_t)] , \tag{12}$$

$$\mathbf{z}_{t+1} = \Pi_{\mathbf{z}\in\mathcal{Z}}[\mathbf{z}_t - \frac{1}{\eta_{\mathbf{z}}} G(\mathbf{z}_t)] . \tag{13}$$

If both Bregman divergences are KL divergence and $\mathcal{X}, \mathcal{Z}$ are unit simplex, Bregman ADM leads to the following alternating multiplicative updates:

$$\mathbf{x}_{t+1,i} = \frac{\mathbf{x}_{t,i} \exp(-\frac{1}{\eta_{\mathbf{x}}} F_i(\mathbf{x}_t))}{\sum_{i=1}^n \mathbf{x}_{t,i} \exp(-\frac{1}{\eta_{\mathbf{x}}} F_i(\mathbf{x}_t))} , \tag{14}$$

$$\mathbf{z}_{t+1,i} = \frac{\mathbf{z}_{t,i} \exp(-\frac{1}{\eta_{\mathbf{z}}} G_i(\mathbf{z}_t))}{\sum_{i=1}^n \mathbf{z}_{t,i} \exp(-\frac{1}{\eta_{\mathbf{z}}} G_i(\mathbf{z}_t))} . \tag{15}$$

If one is quadratic function and the other is KL-divergence, we have alternating additive-multiplicative updates. The convergence rate of Bregman ADMs will be provided in the full paper.

3 Minimizing over Doubly Stochastic Matrices

In this section, as an illustrative example, we consider the problem of minimizing a loss function of a doubly stochastic matrix, which has been studied in spectral clustering [15] and learning permutations [9]. The class of $n \times n$ doubly stochastic matrices is a convex polytope known as the Birkhoff polytope $\mathbf{B}_n$. In particular, we consider the following problem:

$$\min f(\mathbf{P}) \quad \text{s.t.} \quad \mathbf{P} \in \mathbf{B}_n , \tag{16}$$

where $\mathbf{B}_n$ denotes the Birkhoff polytope such that $\mathbf{P} \geq 0, \mathbf{e}^T\mathbf{P} = \mathbf{e}, \mathbf{P}\mathbf{e} = \mathbf{e}$ and $\mathbf{e}$ is a column vector of ones. To solve this particular convex minimization problem, we can use MDA which has the following update:

$$\mathbf{P}^{t+1} = \text{argmin}_{\mathbf{P} \in \mathbf{B}_n} \langle f'(\mathbf{P}^t), \mathbf{P} \rangle + \eta B_\phi(\mathbf{P}, \mathbf{P}^t) . \tag{17}$$

In (17), simply choosing a Bregman divergence does not yield efficient projection onto the Birkhoff polytope. Since $\mathbf{B}_n$ contains the structure of unit simplex ($\mathbf{P} \geq 0, \mathbf{e}^T\mathbf{P} = \mathbf{e}$), we use KL divergence in (17) which yields a multiplicative update. As a result, MDA leads to a double-loop algorithm which has the following two steps:

$$\mathbf{P}_{ij}^{t+\frac{1}{2}} = \mathbf{P}_{ij}^t \exp(-\frac{1}{\eta} L_{ij}) , \tag{18}$$

$$\mathbf{P}^{t+1} = \Pi_{\mathbf{B}_n}(\mathbf{P}^{t+\frac{1}{2}}) . \tag{19}$$

where $\mathbf{L} = f'(\mathbf{P}^t)$ and $\Pi_{\mathbf{B}_n}$ denotes the projection back onto Birkhoff polytope which can be solved using Sinkhorn algorithm [13, 9].

The projection in (19) normalizes column and row to 1 repeatedly and alternatively until convergence. We now show this iterative step can be simply replaced by two additional $O(n^2)$ steps. We split $\mathbf{B}_n$ into an unit simplex $\mathbf{B}_n^c = \{\mathbf{P}_c | \mathbf{P}_c \geq 0, \mathbf{e}^T\mathbf{P}_c = \mathbf{e}\}$ and an equality constraint $\mathbf{P}_r\mathbf{e} = \mathbf{e}$. (16) can be rewritten in the ADM form:

$$\min \quad f(\mathbf{P}_\mathbf{c}) \quad \text{s.t.} \quad \mathbf{P}_\mathbf{c} \in \mathbf{B}_n^c, \mathbf{P}_\mathbf{r}\mathbf{e} = \mathbf{e}, \mathbf{P}_\mathbf{c} = \mathbf{P}_\mathbf{r} . \tag{20}$$

We use updates (10) and (8). Let the Bregman divergence be KL divergence in (10), we have

$$\mathbf{P}_\mathbf{c}^{t+1} = \text{argmin}_{\mathbf{P}_\mathbf{c} \in \mathbf{B}_n^c} \langle A, \mathbf{P}_\mathbf{c} \rangle + \eta KL(\mathbf{P}_\mathbf{c}, \mathbf{P}_\mathbf{c}^t) , \tag{21}$$

$$\mathbf{P}_\mathbf{r}^{t+1} = \text{argmin}_{\mathbf{P}_\mathbf{r}\mathbf{e} = \mathbf{e}} \langle \mathbf{Q}^t, -\mathbf{P}_\mathbf{r} \rangle + \frac{\rho}{2} \|\mathbf{P}_\mathbf{c}^{t+1} - \mathbf{P}_\mathbf{r}\|_2^2 , \tag{22}$$

$$\mathbf{Q}^{t+1} = \mathbf{Q}^t + \tau\rho(\mathbf{P}_\mathbf{c}^{t+1} - \mathbf{P}_\mathbf{r}^{t+1}) . \tag{23}$$

where $A = \mathbf{L} + \mathbf{Q}^t + \rho(\mathbf{P}_\mathbf{c}^t - \mathbf{P}_\mathbf{r}^t)$ and $\mathbf{L} = f'(\mathbf{P}_\mathbf{c}^t)$. (21) yields a multiplicative update given in (24).

We now show (22) has a closed-form solution. The Lagrangian for (22) is

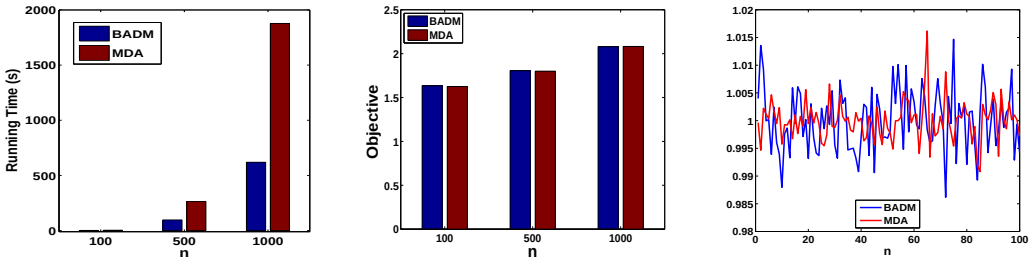
$$L(\mathbf{P}_\mathbf{r}, \mathbf{R}) = \langle \mathbf{Q}^t, -\mathbf{P}_\mathbf{r} \rangle + \frac{\rho}{2} \|\mathbf{P}_\mathbf{c}^{t+1} - \mathbf{P}_\mathbf{r}\|_2^2 + \langle \mathbf{R}, \mathbf{P}_\mathbf{r}\mathbf{e} - \mathbf{e} \rangle .$$

Setting the derivative with respect to $\mathbf{P}_\mathbf{r}$ to zero gives $\mathbf{P}_\mathbf{r} = \mathbf{M} - \mathbf{R}\mathbf{e}^T/\rho$, where $\mathbf{M} = \mathbf{P}_\mathbf{c}^{t+1} + \mathbf{Q}^t/\rho$. Multiplying both sides by $\mathbf{e}$ gives $\mathbf{R} = (\rho(\mathbf{I}_n - \mathbf{M})\mathbf{e}))/n$. Substituting $\mathbf{R}$ back yields (25).

Overall, Bregman ADM yields a single-loop algorithm which has the following updates:

$$\mathbf{P}_{\mathbf{c},ij}^{t+1} = \frac{\mathbf{P}_{\mathbf{c},ij}^t \exp(-\frac{A_{ij}}{\eta})}{\sum_{i=1}^n \mathbf{P}_{\mathbf{c},ij}^t \exp(-\frac{A_{ij}}{\eta})} , \tag{24}$$

$$\mathbf{P}_\mathbf{r}^{t+1} = \mathbf{M} - (\mathbf{I}_n - \mathbf{M})\mathbf{e}\mathbf{e}^T/n , \tag{25}$$



(a) (b) (c)

Figure 1: (a): Running time. (b): the objective value. (c): The sum of row of doubly stochastic matrix for $n = 100$. BADM runs much faster than MDA without the loss of performance.

where $\mathbf{M} = \mathbf{P}_{\mathbf{c}}^{t+1} + \mathbf{Q}^t/\rho$. The three updates can be done in $O(n^2)$.

The following experiment compares BADM and MDA in minimizing a linear function over doubly stochastic matrix. Let $f(\mathbf{P}) = \mathrm{Tr}(\mathbf{L}^T\mathbf{P})$, where $\mathbf{L} \in \mathbb{R}^{n\times n}$ is randomly generated from uniform distribution. We set $\eta = 1$ in MDA and $\rho = 0.5, \eta = 1, \tau = 1$ in BADM. Both algorithms are run 20 times for $n = 100, 500, 1000$ and the average results are reported. The running time is plotted in Figure 1(a) and objective value is plotted in Figure 1(b). In both BADM and MDA, the sum of column of doubly stochastic matrix is always equal to 1. We plot the sum of row of doubly stochastic matrix for $n = 100$ in Figure 1(c), which shows the matrices in BADM and MDA are almost row stochastic. BDAM runs much faster than MDA while maintaining the same performance as MDA.

Acknowledgment

This research was supported by NSF grants IIS-0916750, IIS-0812183, IIS-0534286, NSF CAREER grant IIS-0953274, and NASA grant NNX08AC36A.

References

[1] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[2] S. Boyd, E. C. N. Parikh, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundation Trends Machine Learning*, 3(1):1–122, 2011.

[3] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bremgan functions. *SIAM Journal on Optimization*, 3:538–543, 1993.

[4] P. Combettes and J. Pesquet. Proximal splitting methods in signal processsing. *ArXiv*, 2009.

[5] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, 2010.

[6] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *JMLR*, 10:2873–2898, 2009.

[7] B. He and X. Yuan. On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. *Preprint*, 2012.

[8] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50:700–709, 2012.

[9] D. Helmhold and M. Warmuth. Learning permutations with exponential weights. *JMLR*, 10:1705–1736, 2009.

[10] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132:1–63, 1997.

[11] Y. Nesterov. Gradient methods for minimizing composite objective function. *Technical Report 76, Center for Operation Research and Economics (CORE), Catholic University of Louvain (UCL)*, 2007.

[12] M. Schmidt and F. B. N. Roux. Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*, 2011.

[13] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21:343–348, 1967.

[14] H. Wang and A. Banerjee. Online alternating direction method. In *ICML*, 2012.

[15] R. Zass and A. Shashua. Doubly stochastic normalization for spectral clustering. In *NIPS*, 2006.