Corrections

Trung Vu
Department of CSEE, University of Maryland, Baltimore County, MD 21250, USA
trungvv@umbc.edu

May 4, 2024

1 Equation (13)

It will be more appropriate to say $\epsilon(\cdot)$ is a similarity measure instead of a distance measure.

2 Equation (16)

In the paper, the augmented Lagrangian is given by

$$\mathcal{L}_n^c(\boldsymbol{W}) \triangleq E\big[\log p(\boldsymbol{w}_n^\top \boldsymbol{x})\big] + \log\big|\boldsymbol{d}_n^\top \boldsymbol{w}_n\big| + \frac{1}{2\gamma_n}\Big(\big(\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}\big)^2 - \mu_n^2\Big), \tag{1}$$

where μ_n is a Lagrangian multiplier and γ_n is a positive scalar learning parameter. $h_n(\boldsymbol{w}_n, \boldsymbol{r}_n)$ corresponds to the inequality constraint $h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) = \rho_n - \big|\boldsymbol{w}_n^\top \boldsymbol{r}_n\big| \leq 0$. Also, we note that the goal here is to maximize $\mathcal{L}_n^c(\boldsymbol{W})$ *w.r.t.* $\boldsymbol{W}$. In the following, we review the augmented Lagrangian method and show that there is a **sign issue** for the Lagrangian term in (1).

2.1 Augmented Lagrangian Method

The augmented Lagrangian method, a.k.a. the method of multipliers, is used to handle the inequality constraints as follows. Consider the general setting of a constrained minimization problem

$$\min f(\boldsymbol{x}) \quad \text{subject to } g_j(\boldsymbol{x}) \leq 0, \text{ for } i = j, \ldots, m, \tag{2}$$

where $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ and $g_j(\cdot) : \mathbb{R}^n \to \mathbb{R}$. Let us define the augmented Lagrangian as

$$\mathcal{L}_\gamma(\boldsymbol{x}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \frac{1}{2\gamma}\sum_{j=1}^{m}\Big(\Big(\max\big(0, \mu_j + \gamma g_j(\boldsymbol{x})\big)\Big)^2 - \mu_j^2\Big), \tag{3}$$

where $\boldsymbol{\mu} \in \mathbb{R}^m$ is the Lagrange multiplier and $\gamma > 0$ is a scalar penalty parameter. The iterative equations to minimize (3) are given by

$$\begin{cases} \boldsymbol{x}^{i+1} = \operatorname{argmin}_{\boldsymbol{x}} \mathcal{L}_\gamma(\boldsymbol{x}, \boldsymbol{\mu}^i) \\ \boldsymbol{\mu}^{i+1} = \max\big(\boldsymbol{0}, \boldsymbol{\mu}^i + \gamma \boldsymbol{g}(\boldsymbol{x}^{i+1})\big) \end{cases}, \tag{4}$$

where the operators in the second update are element-wise. It can be shown [1] that for sufficiently large γ, the solution of (3) coincides with the solution of (2).

2.2 Correction of the Sign Issue in Equation (16)

Comparing the minimization in (3) versus the maximization in (1), we see that the sign of $\gamma_n > 0$ is incorrect in (1). If one would like to maximize $\mathcal{L}_n^c(\boldsymbol{W})$, the augmented Lagrangian function should be defined as

$$\max \mathcal{L}_n^c(\boldsymbol{W}) \triangleq E\big[\log p(\boldsymbol{w}_n^\top \boldsymbol{x})\big] + \log\big|\boldsymbol{d}_n^\top \boldsymbol{w}_n\big| - \frac{1}{2\gamma_n}\Big(\big(\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}\big)^2 - \mu_n^2\Big). \tag{5}$$

Alternatively, it is more common to consider the minimization (instead of maximization) by changing the size of the IVA cost term and then proceed with the standard augmented Lagrangian method:

$$\min \mathcal{L}_n^c(\boldsymbol{W}) \triangleq -E\big[\log p(\boldsymbol{w}_n^\top \boldsymbol{x})\big] - \log\big|\boldsymbol{d}_n^\top \boldsymbol{w}_n\big| + \frac{1}{2\gamma_n}\Big(\big(\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}\big)^2 - \mu_n^2\Big). \tag{6}$$

3 Equation (17)

In the paper, the gradient of $\mathcal{L}_n^c(\boldsymbol{W})$ *w.r.t.* $\boldsymbol{w}_n$ is given by

$$\frac{\partial \mathcal{L}_n^c}{\partial \boldsymbol{w}_n} = E\big[f_n(\boldsymbol{w}_n^\top \boldsymbol{x})\boldsymbol{x}^\top\big] + \frac{\boldsymbol{d}_n^\top}{\boldsymbol{d}_n^\top \boldsymbol{w}_n} + h_n'(\boldsymbol{w}_n, \boldsymbol{r}_n)\mu_n \boldsymbol{r}_n^\top, \tag{7}$$

where $h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) = \rho_n - \big|\boldsymbol{w}_n^\top \boldsymbol{r}_n\big|$ and h_n' is the derivative of h_n *w.r.t.* $(\boldsymbol{w}_n^\top \boldsymbol{r}_n)$. Ignoring the sign issue mentioned in the previous section and the transposition of the column vector to the row vector, we focus on the derivation of the gradient of the Lagrangian term:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}_n}\left(\frac{1}{2\gamma_n}\Big(\big(\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}\big)^2 - \mu_n^2\Big)\right) &= \frac{1}{2\gamma_n}\frac{\partial}{\partial \boldsymbol{w}_n}\big(\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}\big)^2 \\
&= \frac{1}{2\gamma_n}\big(2\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}\big)\frac{\partial}{\partial \boldsymbol{w}_n}\big(\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}\big) \\
&= \frac{\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}}{\gamma_n}\mathbb{I}_{\gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n > 0}\frac{\partial}{\partial \boldsymbol{w}_n}\big(\gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\big) \qquad \text{(since } (\partial \max\{0, x\}/\partial x = \mathbb{I}_{x>0}) \\
&= \frac{\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}}{\gamma_n}\gamma_n \frac{\partial h_n(\boldsymbol{w}_n, \boldsymbol{r}_n)}{\partial \boldsymbol{w}_n} \qquad \text{(absorbing } \mathbb{I}_{\gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n > 0} \text{ into the max)} \\
&= \max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}\frac{\partial h_n(\boldsymbol{w}_n, \boldsymbol{r}_n)}{\partial(\boldsymbol{w}_n^\top \boldsymbol{r}_n)}\frac{\partial(\boldsymbol{w}_n^\top \boldsymbol{r}_n)}{\partial \boldsymbol{w}_n} \\
&= \max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\} h'(\boldsymbol{w}_n, \boldsymbol{r}_n)\boldsymbol{r}_n.
\end{aligned} \tag{8}$$

Note that the key difference between (7) and (8) is the term μ_n is replaced by its updated value in the next iteration $\max\{0, \gamma_n h_n(\boldsymbol{w}_n, \boldsymbol{r}_n) + \mu_n\}$ (see 4). This typo, however, does not affect the correctness of the implementation since μ_n is updated before $\Delta \boldsymbol{w}_n$ (see Steps 7 and 8 of Algorithm 2).

References

[1] Dimitri P Bertsekas, *Constrained optimization and Lagrange multiplier methods*, Academic press, 2014.