An introduction to ADMM

November 22, 2019

ToC

Dual Ascent

Consider equality-constrianed convex optimization problem

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t.}\ \ Ax = b \end{aligned} \tag{1.1}$$

Lagrangian for this problem

$$L(x, y) = f(x) + y^T(Ax - b) \tag{1.2}$$

Dual Ascent

Dual Ascent

$$x^{k+1} := \underset{x}{\operatorname{argmin}} f(x) + (y^k)^T(Ax - b) \tag{1.3}$$

$$y^{k+1} := y^k + \alpha^k \left(Ax^{k+1} - b\right) \tag{1.4}$$

If f is a nonzero affine function, then the x-update (1.3) may fail

Dual Decomposition

$$f(x) = \sum_{i=1}^{N} f_i(x_i), A = [A_1 \cdots A_N]$$

$$L(x, y) = \sum_{i=1}^{N} L_i(x_i, y) = \sum_{i=1}^{N} \left(f_i(x_i) + y^T A_i x_i - (1/N) y^T b\right)$$

Dual Ascent

$$\text{(Broadcast) } x_i^{k+1} := \underset{x_i}{\operatorname{argmin}} L_i\left(x_i, y^k\right) \tag{1.5}$$

$$\text{(Gather) } y^{k+1} := y^k + \alpha^k\left(Ax^{k+1} - b\right) \tag{1.6}$$

Augmented Lagrangians

The augmented lagrangian for (1.1) is

$$L_\rho(x,y) = f(x) + y^T(Ax - b) + (\rho/2)\|Ax - b\|_2^2 \tag{2.1}$$

It's the lagrangian for the following minimization problem which is equivalent to the original problem (1.1)

$$\begin{aligned} \min\ & f(x) + (\rho/2)\|Ax - b\|_2^2 \\ \text{s.t.}\ & Ax = b \end{aligned} \tag{2.2}$$

Method of Multipliers

Method of Multipliers

$$x^{k+1} := \underset{x}{\operatorname{argmin}} L_\rho \left(x, y^k\right) \tag{2.3}$$

$$y^{k+1} := y^k + \rho \left(Ax^{k+1} - b\right) \tag{2.4}$$

If f is a nonzero affine function, then the x-update (2.3) doesn't fail

Method of Multipliers

KKT condition for (1.1) is

$$Ax^{\star} - b = 0, \quad \nabla f(x^{\star}) + A^T y^{\star} = 0 \tag{2.5}$$

Choosing step size ρ in (2.4) leads to

$$\begin{aligned} 0 &= \nabla_x L_\rho\left(x^{k+1}, y^k\right) \\ &= \nabla_x f\left(x^{k+1}\right) + A^T\left(y^k + \rho\left(Ax^{k+1} - b\right)\right) \\ &= \nabla_x f\left(x^{k+1}\right) + A^T y^{k+1} \end{aligned}$$

So (x^{k+1}, y^{k+1}) is dual feasible and if $Ax^{k+1} - b$ converges to zero, yielding optimality

Method of Multipliers

- Problem: When f is separable, the augmented Lagrangian L_ρ is not separable
- Solution: Alternating Direction Method of Multipliers

Alternating Direction Method of Multipliers

Consider equality-constrianed convex optimization problem

$$\begin{aligned} \min\ & f(x) + g(z) \\ \text{s.t.}\ & Ax + Bz = c \end{aligned} \tag{3.1}$$

The augmented Lagrangian is

$$L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + (\rho/2)\|Ax + Bz - c\|_2^2 \tag{3.2}$$

Alternating Direction Method of Multipliers

ADMM

$$x^{k+1} := \underset{x}{\operatorname{argmin}} L_\rho\left(x, z^k, y^k\right) \tag{3.3}$$

$$z^{k+1} := \underset{z}{\operatorname{argmin}} L_\rho\left(x^{k+1}, z, y^k\right) \tag{3.4}$$

$$y^{k+1} := y^k + \rho\left(Ax^{k+1} + Bz^{k+1} - c\right) \tag{3.5}$$

Method of Multipliers

$$\left(x^{k+1}, z^{k+1}\right) := \underset{x,z}{\operatorname{argmin}} L_\rho\left(x, z, y^k\right) \tag{3.6}$$

$$y^{k+1} := y^k + \rho\left(Ax^{k+1} + Bz^{k+1} - c\right) \tag{3.7}$$

Let $u = (1/\rho)y$, this leads to the scaled form of ADMM

Scaled Form of ADMM

$$x^{k+1} := \underset{x}{\operatorname{argmin}} \left(f(x) + (\rho/2) \left\| Ax + Bz^k - c + u^k \right\|_2^2 \right) \tag{3.8}$$

$$z^{k+1} := \underset{z}{\operatorname{argmin}} \left(g(z) + (\rho/2) \left\| Ax^{k+1} + Bz - c + u^k \right\|_2^2 \right) \tag{3.9}$$

$$u^{k+1} := u^k + Ax^{k+1} + Bz^{k+1} - c \tag{3.10}$$

Consensus optimization

$$\min \; f(x) = \sum_{i=1}^{N} f_i(x) \Leftrightarrow \begin{aligned} \min \; & f(x) = \sum_{i=1}^{N} f_i(x_i) \\ \text{s.t.} \; & x_i - z = 0 \end{aligned} \tag{3.11}$$

ADMM

$$\text{(Broadcast) } x_i^{k+1} := \underset{x_i}{\operatorname{argmin}} \left(f_i(x_i) + (\rho/2) \left\| x_i - z^k + u_i^k \right\|_2^2 \right) \tag{3.12}$$

$$\text{(Gather) } z^{k+1} := (\rho/2) \sum_{i=1}^{N} \| x_i^{k+1} - z + u_i^k \|_2^2 \tag{3.13}$$

$$\text{(Broadcast) } u_i^{k+1} := u_i^k + \left(x_i^{k+1} - z^{k+1} \right) \tag{3.14}$$

Remark:ADMM is suitable for distributed optimization

Consensus optimization

General form consensus optimization problem

$$\begin{aligned} \min \ & f(x) = \sum_{i=1}^{N} f_i(x_i) + g(z) \\ \text{s.t.} \ & A_i x_i + B_i z = c_i \end{aligned} \tag{3.15}$$

ADMM

$$\text{(Broadcast) } x_i^{k+1} := \underset{x_i}{\operatorname{argmin}} \left(f_i(x_i) + (\rho/2) \left\| A_i x_i + B_i z^k - c_i + u_i^k \right\|_2^2 \right) \tag{3.16}$$

$$\text{(Gather) } z^{k+1} := g(z) + (\rho/2) \sum_{i=1}^{N} \| A_i x_i^{k+1} + B_i z - c_i + u_i^k \|_2^2 \tag{3.17}$$

$$\text{(Broadcast) } u_i^{k+1} := u_i^k + \left(A_i x_i^{k+1} + B_i z^{k+1} - c_i \right) \tag{3.18}$$

General L_1 Regularized Loss Minimization

$$\min\ f(x) + \lambda\|x\|_1 \Leftrightarrow \begin{array}{l} \min\ f(x) + \lambda\|z\|_1 \\ \text{s.t.}\ \ x - z = 0 \end{array} \tag{3.19}$$

ADMM

$$x^{k+1} := \underset{x}{\operatorname{argmin}} \left(f(x) + (\rho/2) \left\| x - z^k + u^k \right\|_2^2 \right) \tag{3.20}$$

$$z^{k+1} := S_{\lambda/\rho} \left(x^{k+1} + u^k \right) \tag{3.21}$$

$$u^{k+1} := u^k + x^{k+1} - z^{k+1} \tag{3.22}$$

Soft thresholding operator is applied component-wise
The L_1 regularization problem is converted to a L_2 regularization problem
A proof is given in Appendix

General L_2 Regularized Loss Minimization

$$\min\ f(x) + \lambda\|x\|_2^2 \Leftrightarrow \begin{array}{l} \min\ f(x) + \lambda\|z\|_2^2 \\ \text{s.t. } x - z = 0 \end{array} \tag{3.23}$$

ADMM

$$x^{k+1} := \underset{x}{\operatorname{argmin}} \left(f(x) + (\rho/2) \left\| x - z^k + u^k \right\|_2^2 \right) \tag{3.24}$$

$$z^{k+1} := \underset{z}{\operatorname{argmin}} \left(\lambda\|z\|_2^2 + (\rho/2) \left\| x^{k+1} - z + u^k \right\|_2^2 \right) \tag{3.25}$$

$$u^{k+1} := u^k + x^{k+1} - z^{k+1} \tag{3.26}$$

The subproblem also involves a L_2 regularization problem, thus in this case, ADMM doesn't help much

Alternating Direction Method of Multipliers

Theorem (Convergence I)

If the following two conditions are satisfied

- **epi**f *and* **epi**g *are both closed nonempty convex sets*
- *The unaugmented Lagrangian* L_0 *has a saddle point*

Then

- $f\left(x^k\right) + g\left(z^k\right) \to p^\star$
- $r^k \triangleq Ax^k + Bz^k - c \to 0$

A proof which comes from [1] is given in ▸ Appendix

Note: Doesn't give the convergence rate

Alternating Direction Method of Multipliers

Optimality conditions

$$Ax^\star + Bz^\star - c = 0 \tag{3.27}$$

$$0 \in \partial f(x^\star) + A^T y^\star \tag{3.28}$$

$$0 \in \partial g(z^\star) + B^T y^\star \tag{3.29}$$

In the previous convergence proof, we have seen

$$0 \in \partial f\left(x^{k+1}\right) + A^T\left(y^{k+1} - \rho B\left(z^{k+1} - z^k\right)\right)$$

$$0 \in \partial g\left(z^{k+1}\right) + B^T y^{k+1}$$

$$\text{(primal residual) } r^{k+1} = Ax^{k+1} + Bz^{k+1} - c$$

$$\text{(dual residual) } s^{k+1} = \rho A^T B\left(z^{k+1} - z^k\right)$$

Alternating Direction Method of Multipliers

Stopping Criteria

$$f\left(x^k\right) + g\left(z^k\right) - p^\star \leq -\left(y^k\right)^T r^k + \left(x^k - x^\star\right)^T s^k$$

$$f\left(x^k\right) + g\left(z^k\right) - p^\star \leq -\left(y^k\right)^T r^k + d\left\|s^k\right\|_2 \leq \left\|y^k\right\|_2 \left\|r^k\right\|_2 + d\left\|s^k\right\|_2$$

$$\left\|r^k\right\|_2 \leq \epsilon^{\mathrm{pri}}, \left\|s^k\right\|_2 \leq \epsilon^{\mathrm{dual}}$$

$$\epsilon^{\mathrm{pri}} = \sqrt{p}\epsilon^{\mathrm{abs}} + \epsilon^{\mathrm{rel}} \max\left\{\left\|Ax^k\right\|_2, \left\|Bz^k\right\|_2, \|c\|_2\right\}$$

$$\epsilon^{\mathrm{dual}} = \sqrt{n}\epsilon^{\mathrm{abs}} + \epsilon^{\mathrm{rel}}\left\|A^T y^k\right\|_2$$

Alternating Direction Method of Multipliers

- In practice, ADMM is slow to converge to high accuracy. But ADMM often converges to modest accuracy within a few tens of iterations, which is sufficient for many large-scale applications [1]
- The ADMM has first been shown to converge with a rate of $O(1/k)$ in the general case of convex non-smooth cost functions in [3] based on a variational inequality approach
- Linear convergence on the other hand has been shown for different applications and assumptions, such as [4],[5]

ADMM-Softmax

Consider softmax classifier

$$h_{\mathbf{W}}(\mathbf{d}) = \frac{\exp(\mathbf{W}\mathbf{d})}{\mathbf{e}_{n_\mathrm{c}}^\top \exp(\mathbf{W}\mathbf{d})} \tag{4.1}$$

where $h_{\mathbf{W}}(\mathbf{d}) \in \mathbb{R}^{n_c}$ is is the predicted class label, $\mathbf{W} \in \mathbb{R}^{n_c \times n_f}$ is a weight matrix, $\mathbf{e}_{n_\mathrm{c}} \in \mathbb{R}^{n_\mathrm{c}}$ is a vector of all ones and the exponential function is applied component-wise

ADMM-Softmax

Expected cross-entropy loss function

$$
\begin{aligned}
\Phi(\mathbf{W}) &= \mathbb{E}\left[-\mathbf{c}^\top \log\left(h_{\mathbf{W}}(\mathbf{d})\right)\right] = \mathbb{E}\left[-\mathbf{c}^\top \log\left(\frac{\exp(\mathbf{Wd})}{\mathbf{e}_{n_c}^\top \exp(\mathbf{Wd})}\right)\right] \\
&= \mathbb{E}\left[-\mathbf{c}^\top \mathbf{Wd} + \left(\mathbf{c}^\top \mathbf{e}_{n_c}\right)\left(\log\left(\mathbf{e}_{n_c}^\top \exp(\mathbf{Wd})\right)\right)\right] \\
&= \mathbb{E}\left[-\mathbf{c}^\top \mathbf{Wd} + \log\left(\mathbf{e}_{n_c}^\top \exp(\mathbf{Wd})\right)\right]
\end{aligned} \tag{4.2}
$$

Sample cross-entropy loss function

$$
\frac{1}{N}\sum_{j=1}^{N}\left[-\mathbf{c}_j^\top \mathbf{Wd}_j + \log\left(\mathbf{e}_{n_c}^\top \exp\left(\mathbf{Wd}_j\right)\right)\right] \tag{4.3}
$$

ADMM-Softmax

$$\min_w \frac{1}{N}\sum_{j=1}^{N}\left[-\mathbf{c}_j^\top \mathbf{W}\mathbf{d}_j + \log\left(\mathbf{e}_{n_c}^\top \exp\left(\mathbf{W}\mathbf{d}_j\right)\right)\right] + \frac{\alpha}{2}\left\|\mathbf{L}\left(\mathbf{W} - \mathbf{W}_{\mathrm{ref}}\right)^\top\right\|_F^2 \tag{4.4}$$

$$\Updownarrow$$

$$\min_{w,z_1,\cdots,z_N} \frac{1}{N}\sum_{j=1}^{N}\left[-\mathbf{c}_j^\top \mathbf{z}_j + \log\left(\mathbf{e}_{n_c}^\top \exp\left(\mathbf{z}_j\right)\right)\right] + \frac{\alpha}{2}\left\|\mathbf{L}\left(\mathbf{W} - \mathbf{W}_{\mathrm{ref}}\right)^\top\right\|_F^2$$
$$\text{s.t. } \mathbf{z}_j = \mathbf{W}\mathbf{d}_j \tag{4.5}$$

ADMM-Softmax

ADMM-Softmax([6])

$$\mathbf{W}^{(k+1)} = \arg\min_{\mathbf{W}} \frac{\rho}{2} \sum_{j=1}^{N} \left(\left\| \mathbf{z}_j^{(k)} - \mathbf{W}\mathbf{d}_j + \mathbf{u}_j^{(k)} \right\|_2^2 \right) + \frac{\alpha}{2} \left\| \mathbf{L} \left(\mathbf{W} - \mathbf{W}_{\mathrm{ref}} \right)^\top \right\|_F^2 \tag{4.6}$$

$$\mathbf{z}_j^{(k+1)} = \arg\min_{\mathbf{z}_j} - \mathbf{c}_j^\top \mathbf{z}_j + \log \left(\mathbf{e}_{n_c}^\top \exp\left(\mathbf{z}_j\right) \right) + \frac{\rho}{2} \left\| \mathbf{z}_j - \mathbf{W}^{(k+1)}\mathbf{d}_j + \mathbf{u}_j^{(k)} \right\|_2^2 \tag{4.7}$$

$$\mathbf{u}_j^{(k+1)} = \mathbf{u}_j^{(k)} + \left(\mathbf{z}_j^{(k+1)} - \mathbf{W}^{(k+1)}\mathbf{d}_j \right) \tag{4.8}$$

ADMM-Softmax

- W-update

$$\mathbf{A}_{\alpha,\rho}\mathbf{W}^\top = \rho\mathbf{D}(\mathbf{Z}+\mathbf{U})^\top + \alpha\mathbf{L}\mathbf{L}^\top\mathbf{W}_{\mathrm{ref}}^\top, \mathbf{A}_{\alpha,\rho} = \rho\mathrm{D}\mathrm{D}^{\boldsymbol{T}} + \alpha\mathbf{L}^{\boldsymbol{T}}\mathbf{L}$$

- Z-update

$$\Psi(\mathbf{z}_j) = -\mathbf{c}_j^\top\mathbf{z}_j + \log\left(\mathbf{e}_{n_\mathrm{c}}^\top \exp(\mathbf{z}_j)\right) + \frac{\rho}{2}\left\|\mathbf{z}_j - \mathbf{z}_{j,\mathrm{ref}}\right\|_2^2$$

$$\nabla_{\mathbf{z}_j}\Psi(\mathbf{z}_j) = -\mathbf{c}_j + \frac{\exp(\mathbf{z}_j)}{\mathbf{e}_{n_\mathrm{c}}^\top \exp(\mathbf{z}_j)} + \rho\left(\mathbf{z}_j - \mathbf{z}_{j,\mathrm{ref}}\right) \in \mathbb{R}^{n_\mathrm{c}}$$

$$\nabla_{\mathbf{z}_j}^2\Psi(\mathbf{z}_j) = \frac{\operatorname{diag}(\exp(\mathbf{z}_j))}{\mathbf{e}_{n_c}^\top \exp(\mathbf{z}_j)} - \frac{\exp(\mathbf{z}_j)\exp(\mathbf{z}_j)^\top}{\left(\mathbf{e}_{n_c}^\top \exp(\mathbf{z}_j)\right)^2} + \rho\operatorname{diag}(\mathbf{z}_j) \in \mathbb{R}^{n_c \times n_c}$$

Stopping Criteria

$$\left\|\mathbf{r}^{(k+1)}\right\|_2 = \sum_{j=1}^{N} \left\|\mathbf{r}_j^{(k+1)}\right\|_2 = \sum_{j=1}^{N} \left\|\mathbf{z}_j^{(k+1)} - \mathbf{W}^{(k+1)}\mathbf{d}_j\right\|_2$$

$$\left\|\mathbf{s}^{(k+1)}\right\|_2 = \sum_{j=1}^{N} \left\|\mathbf{s}_j^{(k+1)}\right\|_2 = \sum_{j=1}^{N} \left\|\rho \operatorname{vec}\left(\left(\mathbf{z}_j^{(k+1)} - \mathbf{z}_j^{(k)}\right) \mathbf{d}_j^\top\right)\right\|_2$$

$$\left\|\mathbf{r}^{(k+1)}\right\|_2 \leq \epsilon_{\mathrm{pri}}^{(k)} = \sqrt{N n_{\mathrm{c}}}\epsilon_{\mathrm{abs}} + \epsilon_{\mathrm{rel}} \max\left\{\left\|\mathbf{z}_1^{(k)}\right\|_2, \ldots, \left\|\mathbf{z}_N^{(k)}\right\|_2, \left\|\mathbf{W}^{(k)}\mathbf{d}_1\right\|_2, \ldots, \left\|\mathbf{W}^{(k)}\mathbf{d}_N\right\|_2\right\}$$

$$\left\|\mathbf{s}^{(k+1)}\right\|_2 \leq \epsilon_{\mathrm{dual}}^{(k)} = \sqrt{N n_{\mathrm{c}}}\epsilon_{\mathrm{abs}} + \epsilon_{\mathrm{rel}} \max\left\{\left\|\mathbf{u}_1^{(k)}\right\|_2, \ldots, \left\|\mathbf{u}_N^{(k)}\right\|_2\right\}$$

Numerical experiments

- Results can be reproduced using the codes provided at https://github.com/swufung/ADMMSoftmax

Figure 4.1: Example of hand-written images obtained from the MNIST data set

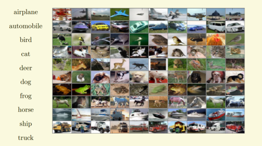

Figure 4.2: Example images for the CIFAR-10 dataset

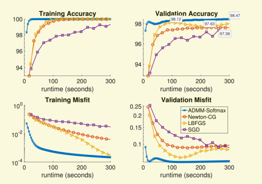

Figure 4.3: MNIST data set

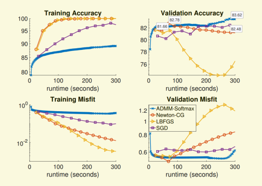

Figure 4.4: CIFAR-10 dataset

Numerical experiments

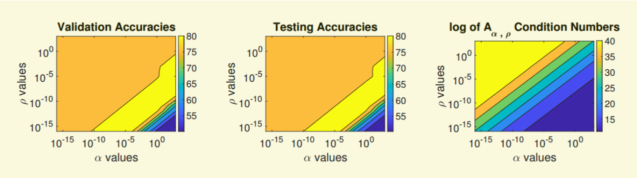

Figure 4.5: Parameter dependence for CIFAR-10 dataset

The validation dataset gives us a good indication of the generalizability of our classifier during the optimization

Appendix

Let $(x^\star, z^\star, y^\star)$ be a saddle point for L_0, $p^\star = f(x^\star) + g(z^\star)$ and define

$$V^k = (1/\rho)\left\|y^k - y^\star\right\|_2^2 + \rho\left\|B\left(z^k - z^\star\right)\right\|_2^2 \tag{5.1}$$

Sketch of Proof

- $V^{k+1} \le V^k - \rho\left\|r^{k+1}\right\|_2^2 - \rho\left\|B\left(z^{k+1} - z^k\right)\right\|_2^2$ (Ineq(1))
- $p^{k+1} - p^\star \le -\left(y^{k+1}\right)^T r^{k+1} - \rho\left(B\left(z^{k+1} - z^k\right)\right)^T\left(-r^{k+1} + B\left(z^{k+1} - z^\star\right)\right)$ (Ineq(2))
- $p^\star - p^{k+1} \le y^{\star T} r^{k+1}$ (Ineq(3))

Then $\rho\sum_{k=0}^{\infty}\left(\left\|r^{k+1}\right\|_2^2 + \left\|B\left(z^{k+1} - z^k\right)\right\|_2^2\right) \le V^0$ leads to the desired convergence

Appendix (Saddle point)

Saddle point

$$L_0(x^\star, z^\star, y) \le L_0(x^\star, z^\star, y^\star) \le L_0(x, z, y^\star) \tag{5.2}$$

$L_0(x^\star, z^\star, y) \le L_0(x^\star, z^\star, y^\star) \Rightarrow Ax^\star + Bz^\star - c = 0$
Thus, $p^\star \le f(x^\star) + g(z^\star) = L_0(x^\star, z^\star, y^\star) = g(y^\star) \le d^\star$
Weak duality $p^\star \ge d^\star$
Thus $p^\star = d^\star = f(x^\star) + g(z^\star) = g(y^\star)$

Appendix (Convergence proof)

Proof of (Ineq(3)).

$$L_0\left(x^\star, z^\star, y^\star\right) \leq L_0\left(x^{k+1}, z^{k+1}, y^\star\right) \Leftrightarrow p^\star \leq p^{k+1} + {y^\star}^T r^{k+1} \tag{5.3}$$

$\square$

Appendix (Convergence proof)

Proof of (Ineq(2)).

$$0 \in \partial L_\rho\left(x^{k+1}, z^k, y^k\right) = \partial f\left(x^{k+1}\right) + A^T y^k + \rho A^T\left(Ax^{k+1} + Bz^k - c\right)$$

$$y^{k+1} = y^k + \rho r^{k+1}$$

$$0 \in \partial f\left(x^{k+1}\right) + A^T\left(y^{k+1} - \rho B\left(z^{k+1} - z^k\right)\right)$$

$$x^{k+1} \text{ minimizes } f(x) + \left(y^{k+1} - \rho B\left(z^{k+1} - z^k\right)\right)^T Ax$$

$$\text{Similarly } z^{k+1} \text{ minimizes } g(z) + y^{(k+1)T} Bz$$

$\square$

Appendix (Convergence proof)

Proof of (Ineq(2)) continue.

$$
\begin{aligned}
& f\left(x^{k+1}\right) + \left(y^{k+1} - \rho B\left(z^{k+1} - z^{k}\right)\right)^{T} A x^{k+1} \\
\leq & f(x^{\star}) + \left(y^{k+1} - \rho B\left(z^{k+1} - z^{k}\right)\right)^{T} A x^{\star}
\end{aligned}
\tag{5.4}
$$

$$
g\left(z^{k+1}\right) + y^{(k+1)T} B z^{k+1} \leq g(z^{\star}) + y^{(k+1)T} B z^{\star} \tag{5.5}
$$

Adding (5.4) and (5.5) leads to (Ineq(2))

$$
\begin{aligned}
p^{k+1} - p^{\star} &\leq -\left(y^{k+1}\right)^{T} r^{k+1} - \rho\left(B\left(z^{k+1} - z^{k}\right)\right)^{T}\left(-r^{k+1} + B\left(z^{k+1} - z^{\star}\right)\right) \\
&= -\left(y^{k+1}\right)^{T} r^{k+1} + \left(x^{k+1} - x^{\star}\right)^{T} s^{k+1} \\
&\text{where } s^{k+1} = \rho A^{T} B\left(z^{k+1} - z^{k}\right)
\end{aligned}
\tag{5.6}
$$

$\square$

Appendix (Convergence proof)

Proof of (Ineq(1)).

Adding (Ineq(2)) and (Ineq(3)) leads to

$$2\left(y^{k+1}-y^{\star}\right)^{T} r^{k+1}-2\rho\left(B\left(z^{k+1}-z^{k}\right)\right)^{T} r^{k+1} +2\rho\left(B\left(z^{k+1}-z^{k}\right)\right)^{T}\left(B\left(z^{k+1}-z^{\star}\right)\right) \leq 0$$

$$y^{k+1}=y^{k}+\rho r^{k+1}$$

$$2\left(y^{k+1}-y^{\star}\right)^{T} r^{k+1}=(1/\rho)\left(\left\|y^{k+1}-y^{\star}\right\|_2^2-\left\|y^{k}-y^{\star}\right\|_2^2\right)+\rho\left\|r^{k+1}\right\|_2^2 \tag{5.7}$$

□

Appendix (Convergence proof)

Proof of (Ineq(1)) continue.

$$z^{k+1} - z^\star = \left(z^{k+1} - z^k\right) + \left(z^k - z^\star\right)$$

$$\rho \left\| r^{k+1} \right\|_2^2 - 2\rho \left(B\left(z^{k+1} - z^k\right)\right)^T r^{k+1}$$

$$+2\rho \left(B\left(z^{k+1} - z^k\right)\right)^T \left(B\left(z^{k+1} - z^\star\right)\right)$$

$$= \rho \left\| r^{k+1} - B\left(z^{k+1} - z^k\right) \right\|_2^2 + \rho \left(\left\| B\left(z^{k+1} - z^\star\right) \right\|_2^2 - \left\| B\left(z^k - z^\star\right) \right\|_2^2 \right) \tag{5.8}$$

$$V^k - V^{k+1} \geq \rho \left\| r^{k+1} - B\left(z^{k+1} - z^k\right) \right\|_2^2$$

$$= \rho \left\| r^{k+1} \right\|_2^2 + \rho \left\| B\left(z^{k+1} - z^k\right) \right\|_2^2 - 2\rho r^{(k+1)T} \left(B\left(z^{k+1} - z^k\right)\right) \tag{5.9}$$

□

Appendix (Convergence proof)

Proof of (Ineq(1)) continue.

$$\text{It suffices to show } 2\rho r^{(k+1)T}\left(B\left(z^{k+1}-z^{k}\right)\right)\leq 0$$

We have seen

$$z^{k+1} \text{ minimizes } g(z)+y^{(k+1)T}Bz$$

$$g\left(z^{k+1}\right)+y^{(k+1)T}Bz^{k+1}\leq g\left(z^{k}\right)+y^{(k+1)T}Bz^{k} \tag{5.10}$$

$$g\left(z^{k}\right)+y^{kT}Bz^{k}\leq g\left(z^{k+1}\right)+y^{kT}Bz^{k+1} \tag{5.11}$$

Adding (5.10) and (5.11) leads to

$$\left(y^{k+1}-y^{k}\right)^{T}\left(B\left(z^{k+1}-z^{k}\right)\right)=\rho r^{(k+1)T}\left(B\left(z^{k+1}-z^{k}\right)\right)\leq 0$$

Go back

□

Appendix (Subgradient)

Subgradient

$$\partial f(x) = \{g | f(y) \geq f(x) + g^T(y - x), \forall y\} \tag{5.12}$$

Example

$$f(x) = |x|$$

$$\partial f(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \\ [-1, 1] & x = 0 \end{cases}$$

Appendix (Subgradient)

- $\partial f(x) + g(x) = \partial f(x) + \partial g(x)$
- If $f \in C^1$, then $\partial f(x) = \{\nabla f(x)\}$

Theorem

$$x^\star \text{ minimizes } f(x) \Leftrightarrow 0 \in \partial f(x^\star)$$

Proof.

$$f(x) \geq f(x^\star) \Leftrightarrow 0 \in \partial f(x^\star)$$

□

Appendix (Subgradient)

$$\min\ f(x) = \lambda|x| + \frac{\rho}{2}(x - v)^2 \tag{5.13}$$

$$\partial f(x) = \begin{cases} \lambda + \rho(x - v) & x > 0 \\ -\lambda + \rho(x - v) & x < 0 \\ [-\lambda - \rho v, \lambda - \rho v] & x = 0 \end{cases}$$

$$x^\star = \begin{cases} v - \frac{\lambda}{\rho} & v > \frac{\lambda}{\rho} \\ 0 & |v| \leq \frac{\lambda}{\rho} \\ v + \frac{\lambda}{\rho} & v < -\frac{\lambda}{\rho} \end{cases}$$

Define soft thresholding operator S

$$S_k(a) = \begin{cases} a - \kappa & a > k \\ 0 & |a| \leq \kappa \\ a + \kappa & a < -k \end{cases}$$

$$\Rightarrow x^\star = S_{\frac{\lambda}{\rho}}(v)$$

Go back

Bibliography I

- Boyd, Stephen, et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers." Foundations and Trends® in Machine learning 3.1 (2011): 1-122.
- Boyd, Stephen, Lin Xiao, and Almir Mutapcic. "Subgradient methods." lecture notes of EE392o, Stanford University, Autumn Quarter 2004 (2003): 2004-2005.
- He, Bingsheng, and Xiaoming Yuan. "On the O(1/n) Convergence Rate of the Douglas–Rachford Alternating Direction Method." SIAM Journal on Numerical Analysis 50.2 (2012): 700-709.
- Hong, Mingyi, and Zhi-Quan Luo. "On the linear convergence of the alternating direction method of multipliers." Mathematical Programming 162.1-2 (2017): 165-199.

Bibliography II

- Shi, Wei, et al. "On the linear convergence of the ADMM in decentralized consensus optimization." IEEE Transactions on Signal Processing 62.7 (2014): 1750-1761.
- Fung, Samy Wu, et al. "ADMM-Softmax: An ADMM Approach for Multinomial Logistic Regression." arXiv preprint arXiv:1901.09450 (2019).
- Bastianello, Nicola. "Distributed Convex Optimisation using the Alternating Direction Method of Multipliers (ADMM) in Lossy Scenarios." (2018).
- http://maths.nju.edu.cn/~hebma/
- https://stanford.edu/~boyd/admm.html