Alternating Direction Method of Multipliers ADMM

Consider a problem of the form:

$$\min_{x,z} \; f(x) + g(z) \;\; \text{subject to} \;\; Ax + Bz = c$$

We define augmented Lagrangian, for a parameter $\rho > 0$,

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

We repeat, for $k = 1, 2, 3, \ldots$

$$\begin{aligned} x^{(k)} &= \operatorname*{argmin}_x \; L_\rho(x, z^{(k-1)}, u^{(k-1)}) \\ z^{(k)} &= \operatorname*{argmin}_z \; L_\rho(x^{(k)}, z, u^{(k-1)}) \\ u^{(k)} &= u^{(k-1)} + \rho(Ax^{(k)} + Bz^{(k)} - c) \end{aligned}$$

ADMM convergence

Under modest assumptions on f, g (these do not require A, B to be full rank), the ADMM iterates satisfy, for any $\rho > 0$:

- Residual convergence: $r^{(k)} = Ax^{(k)} - Bz^{(k)} - c \to 0$ as $k \to \infty$, i.e., primal iterates approach feasibility
- Objective convergence: $f(x^{(k)}) + g(z^{(k)}) \to f^\star + g^\star$, where $f^\star + g^\star$ is the optimal primal objective value
- Dual convergence: $u^{(k)} \to u^\star$, where $u^\star$ is a dual solution

For details, see Boyd et al. (2010). Note that we do not generically get primal convergence, but this is true under more assumptions

Convergence rate: not known in general, theory is currently being developed, e.g., in Hong and Luo (2012), Deng and Yin (2012), Iutzeler et al. (2014), Nishihara et al. (2015). Roughly, it behaves like a first-order method (or a bit faster)

ADMM scaled form

It is often easier to express the ADMM algorithm in a scaled form, where we replace the dual variable u by a scaled variable $w = u/\rho$. In this parametrization, the ADMM steps are:

$$\begin{aligned}
x^{(k)} &= \underset{x}{\operatorname{argmin}}\ f(x) + \frac{\rho}{2}\|Ax + Bz^{(k-1)} - c + w^{(k-1)}\|_2^2 \\
z^{(k)} &= \underset{z}{\operatorname{argmin}}\ g(z) + \frac{\rho}{2}\|Ax^{(k)} + Bz - c + w^{(k-1)}\|_2^2 \\
w^{(k)} &= w^{(k-1)} + Ax^{(k)} + Bz^{(k)} - c
\end{aligned}$$

Note that here the kth iterate $w^{(k)}$ is just given by a running sum of residuals:

$$w^{(k)} = w^{(0)} + \sum_{i=1}^{k} \left(Ax^{(i)} + Bz^{(i)} - c\right)$$

Connection to prox operators

Consider

$$\min_{x} \; f(x) + g(x) \quad \Longleftrightarrow \quad \min_{x,z} \; f(x) + g(z) \;\; \text{subject to} \;\; x = z$$

ADMM steps

$$\begin{aligned} x^{(k)} &= \operatorname{prox}_{f,1/\rho}(z^{(k-1)} - w^{(k-1)}) \\ z^{(k)} &= \operatorname{prox}_{g,1/\rho}(x^{(k)} + w^{(k-1)}) \\ w^{(k)} &= w^{(k-1)} + x^{(k)} - z^{(k)} \end{aligned}$$

where $\operatorname{prox}_{f,1/\rho}$ is the proximal operator for f at parameter $1/\rho$, and similarly for $\operatorname{prox}_{g,1/\rho}$

In general, the update for block of variables reduces to prox update whenever the corresponding linear transformation is the identity

Example: Lasso regression

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, recall the lasso problem:

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

We can rewrite this as:

$$\min_{\beta,\alpha} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\alpha\|_1 \quad \text{subject to} \quad \beta - \alpha = 0$$

ADMM gives us a simple algorithm:

$$\begin{aligned}
\beta^{(k)} &= (X^T X + \rho I)^{-1}(X^T y + \rho(\alpha^{(k-1)} - w^{(k-1)})) \\
\alpha^{(k)} &= S_{\lambda/\rho}(\beta^{(k)} + w^{(k-1)}) \\
w^{(k)} &= w^{(k-1)} + \beta^{(k)} - \alpha^{(k)}
\end{aligned}$$

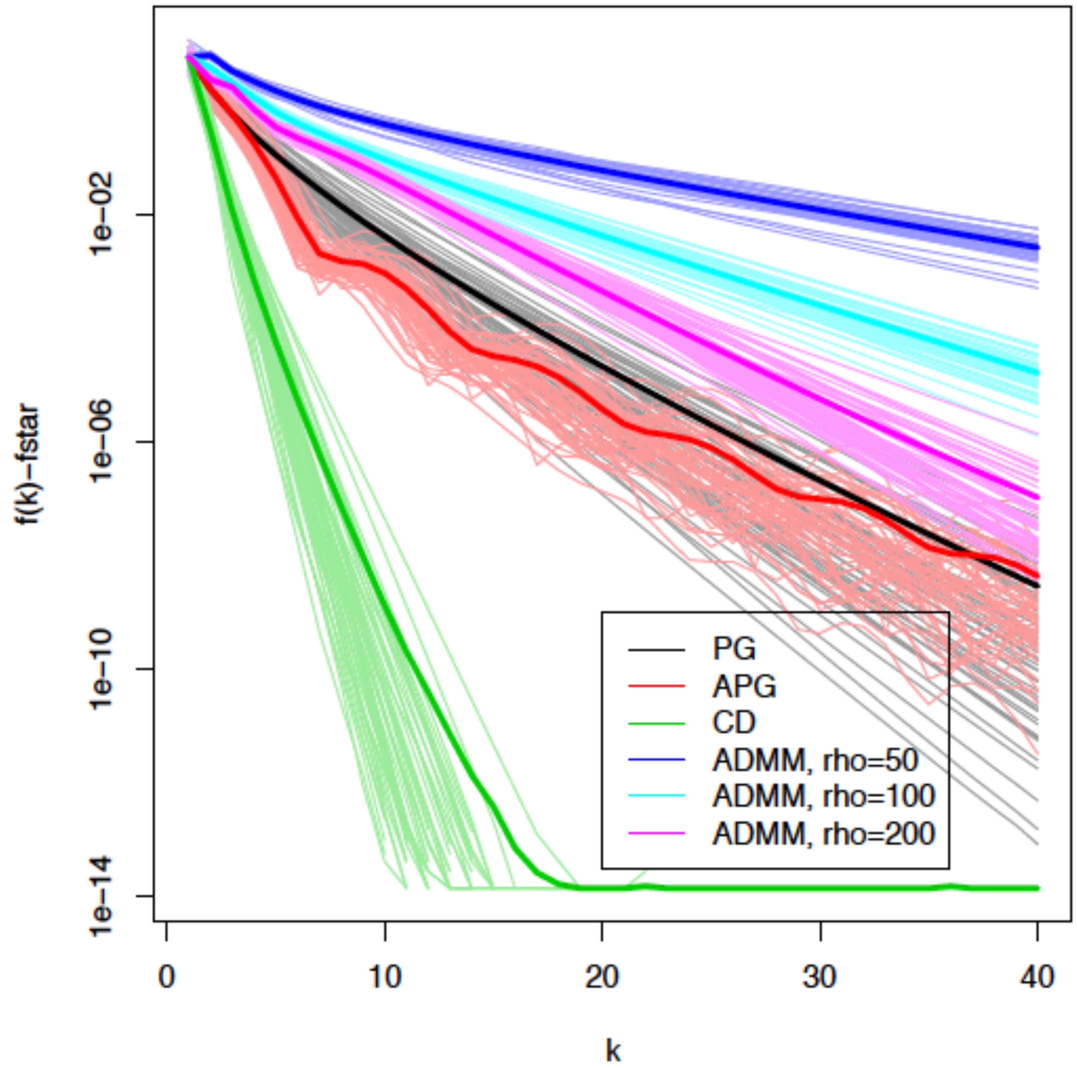
Notes:

- The matrix $X^TX + \rho I$ is always invertible, regardless of X
- If we compute a factorization (say Cholesky) in $O(p^3)$ flops, then each β update takes $O(p^2)$ flops
- The α update applies the soft-thresolding operator S_t, which recall is defined as

$$[S_t(x)]_j = \begin{cases} x_j - t & x > t \\ 0 & -t \leq x \leq t \\ x_j + t & x < -t \end{cases}, \quad j = 1, \ldots p$$

- ADMM steps are "almost" like repeated soft-thresholding of ridge regression coefficients

Comparison of various algorithms for lasso regression: 50 instances with $n = 100$, $p = 20$



Practical issues

In practice, ADMM usually obtains a relatively accurate solution in a handful of iterations, but it requires a large number of iterations for a highly accurate solution (generally behaves like a first-order method)

Choice of ρ can greatly influence practical convergence of ADMM:

- ρ too large → not enough emphasis on minimizing $f + g$
- ρ too small → not enough emphasis on feasibility

Boyd et al. (2010) give a strategy for varying ρ; can be useful, but does not have convergence guarantees

Like deriving duals, transforming a problem into one that ADMM can handle is sometimes a bit subtle, since different forms can lead to different algorithms

Example: Sparse Subspace estimation

Given $S \in \mathbb{S}_p$ (typically $S \succeq 0$ is a covariance matrix), consider the sparse subspace estimation problem (Vu et al., 2013):

$$\max_Y \ \mathrm{tr}(SY) - \lambda \|Y\|_1 \ \text{ subject to } \ Y \in \mathcal{F}_k$$

where $\mathcal{F}_k$ is the Fantope of order k, namely

$$\mathcal{F}_k = \{Y \in \mathbb{S}^p : 0 \preceq Y \preceq I, \ \mathrm{tr}(Y) = k\}$$

Note that when $\lambda = 0$, the above problem is equivalent to ordinary principal component analysis (PCA)

This above is an SDP and in principle solveable with interior point methods, though these can be complicated to implement and quite slow for large problem sizes

We rewrite the problem as:

$$\min_{Y,Z} -\mathrm{tr}(SY) + I_{\mathcal{F}_k}(Y) + \lambda\|Z\|_1 \text{ subject to } Y = Z$$

ADMM steps are:

$$\begin{aligned} Y^{(k)} &= P_{\mathcal{F}_k}(Z^{(k-1)} - W^{(k-1)} + S/\rho) \\ Z^{(k)} &= S_{\lambda/\rho}(Y^{(k)} + W^{(k-1)}) \\ W^{(k)} &= W^{(k-1)} + Y^{(k)} - Z^{(k)} \end{aligned}$$

Here $P_{\mathcal{F}_k}$ is Fantope projection operator, computed by clipping the eigendecomposition $A = U\Sigma U^T$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$:

$$P_{\mathcal{F}_k}(A) = U\Sigma_\theta U^T, \quad \Sigma_\theta = \mathrm{diag}(\sigma_1(\theta), \ldots, \sigma_p(\theta))$$

where each $\sigma_i(\theta) = \min\{\max\{\sigma_i - \theta, 0\}, 1\}$, and $\sum_{i=1}^p \sigma_i(\theta) = k$

Example: Low Rank + Sparse Decomposition

Given $M \in \mathbb{R}^{n \times m}$, consider the sparse plus low rank decomposition problem (Candes et al., 2009):

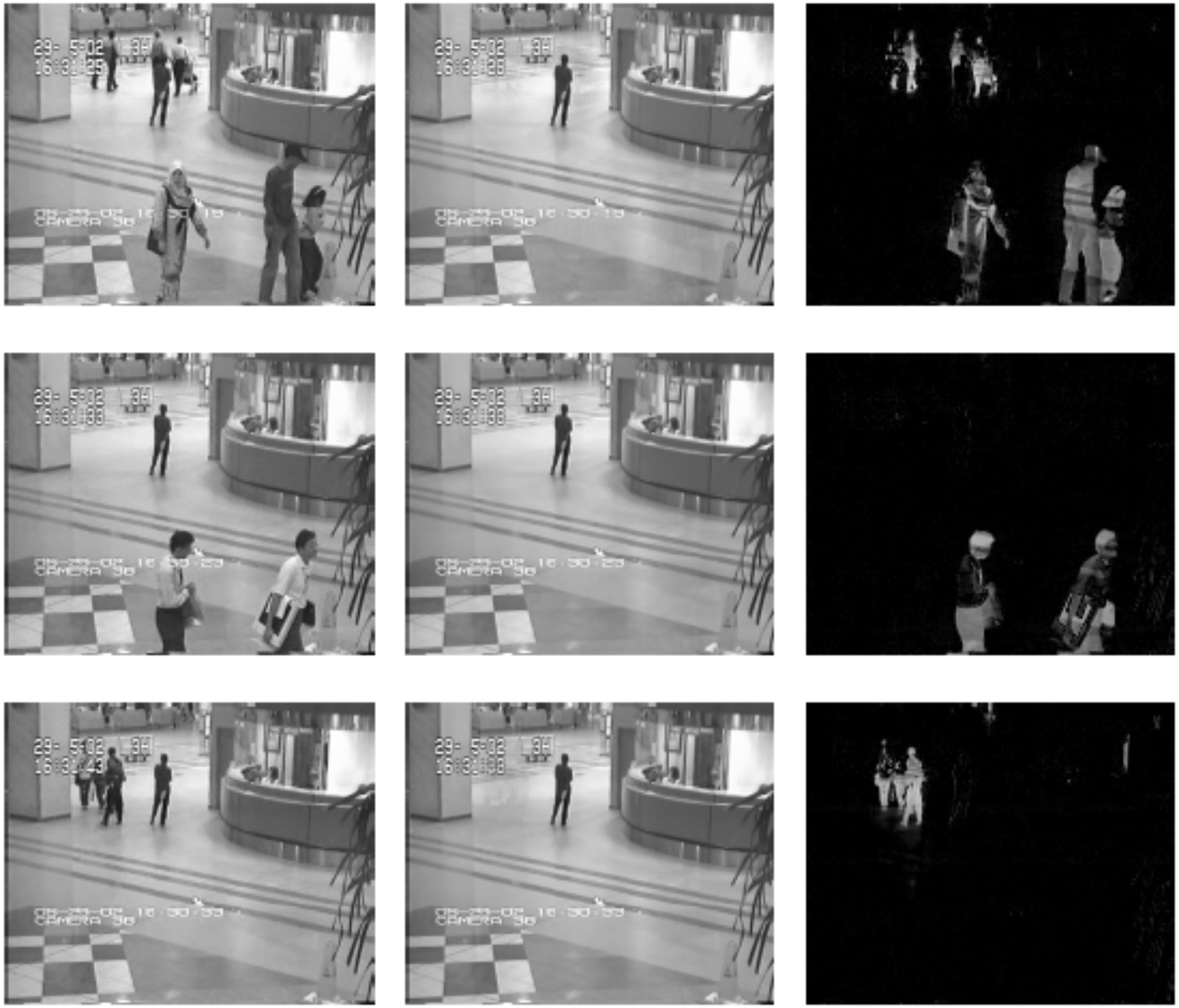
$$\begin{array}{ll} \min_{L,S} & \|L\|_{\mathrm{tr}} + \lambda \|S\|_1 \\ \text{subject to} & L + S = M \end{array}$$

ADMM steps:

$$\begin{aligned} L^{(k)} &= S^{\mathrm{tr}}_{1/\rho}(M - S^{(k-1)} + W^{(k-1)}) \\ S^{(k)} &= S^{\ell_1}_{\lambda/\rho}(M - L^{(k)} + W^{(k-1)}) \\ W^{(k)} &= W^{(k-1)} + M - L^{(k)} - S^{(k)} \end{aligned}$$

where, to distinguish them, we use $S^{\mathrm{tr}}_{\lambda/\rho}$ for matrix soft-thresolding and $S^{\ell_1}_{\lambda/\rho}$ for elementwise soft-thresolding

Example from Candes et al. (2009):



(a) Original frames (b) Low-rank $\hat{L}$ (c) Sparse $\hat{S}$

Faster Convergence?

ADMM can exhibit much faster convergence than usual, when we parametrize subproblems in a "special way"

- ADMM updates relate closely to block coordinate descent, in which we optimize a criterion in an alternating fashion across blocks of variables
- With this in mind, get fastest convergence when minimizing over blocks of variables leads to updates in nearly orthogonal directions
- Suggests we should design ADMM form (auxiliary constraints) so that primal updates de-correlate as best as possible
- This is done in, e.g., Ramdas and Tibshirani (2014), Wytock et al. (2014), Barbero and Sra (2014)

References

- A. Barbero and S. Sra (2014), “Modular proximal optimization for multidimensional total-variation regularization”
- S. Boyd and N. Parikh and E. Chu and B. Peleato and J. Eckstein (2010), “Distributed optimization and statistical learning via the alternating direction method of multipliers”
- E. Candes and X. Li and Y. Ma and J. Wright (2009), “Robust principal component analysis?”
- N. Parikh and S. Boyd (2013), “Proximal algorithms”
- A. Ramdas and R. Tibshirani (2014), “Fast and flexible ADMM algorithms for trend filtering”
- V. Vu and J. Cho and J. Lei and K. Rohe (2013),“Fantope projection and selection: a near-optimal convex relaxation of sparse PCA”
- M. Wytock and S. Sra. and Z. Kolter (2014), “Fast Newton methods for the group fused lasso”