Adaptive ADMM with Spectral Penalty Parameter Selection

Zheng Xu\textsuperscript{1}, Mário A. T. Figueiredo\textsuperscript{2}, Tom Goldstein\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, University of Maryland, College Park, MD
\textsuperscript{2}Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal

April, 2017
Outline

- Constrained problem and alternating direction method of multipliers (ADMM)
- Penalty parameter: crucial for practical convergence
- Spectral penalty parameter selection: fast and fully automated
- Numerical results on various applications and datasets
Constrained problem and ADMM

Constrained problem

\[
\min_{u,v} \ H(u) + G(v) \quad \text{subject to} \quad Au + Bv = b.
\]
Typical applications

- Sparse linear regression (Elastic net regularizer)
  \[
  \min_x \frac{1}{2} \| Dx - c \|_2^2 + \rho_1 \| x \|_1 + \frac{\rho_2}{2} \| x \|_2^2
  \]

- Low rank problem (Nuclear norm regularizer)
Typical applications

- Sparse linear regression (Elastic net regularizer)
- Low rank problem (Nuclear norm regularizer)
- Basis pursuit
- Semidefinite programming
- Dual of SVM / quadratic programming
Typical applications

- Consensus problem for distributed computing

\[
\min_{x_i, z} \sum_{i=1}^{N} f_i(x_i) + g(z) \quad \text{s.t.} \quad x_i - z = 0, \ i = 1, \ldots, N.
\]
Typical applications

- Consensus problem for distributed computing

\[
\min_{x_i, z} \sum_{i=1}^{N} f(x_i) + g(z) \quad \text{s.t.} \quad x_i - z = 0, \; i = 1, \ldots, N.
\]

- More applications: neural networks, tensor decomposition, phase retrieval, robust PCA, TV image problem [Taylor et al., 2016, Xu et al., 2016a,b, 2017]
Constrained problem and ADMM

Constrained problem

\[
\min_{u,v} \ H(u) + G(v) \ \text{subject to} \ Au + Bv = b.
\]

Saddle point problem with augmented Lagrangian

\[
\max_{\lambda} \min_{u,v} \ H(u) + G(v) + \langle \lambda, b - Au - Bv \rangle + \frac{\tau}{2} \|b - Au - Bv\|^2
\]
Constrained problem and ADMM

Constrained problem

\[
\min_{u,v} \ H(u) + G(v) \quad \text{subject to} \quad Au + Bv = b.
\]

Saddle point problem with augmented Lagrangian

\[
\max_{\lambda} \min_{u,v} \ H(u) + G(v) + \langle \lambda, b - Au - Bv \rangle + \frac{\tau}{2} \| b - Au - Bv \|^2
\]

Alternating direction method of multipliers (ADMM)

\[
\begin{align*}
    u_{k+1} &= \arg \min_u H(u) + \langle \lambda_k, -Au \rangle + \frac{\tau}{2} \| b - Au - Bv_k \|^2 \\
v_{k+1} &= \arg \min_v G(v) + \langle \lambda_k, -Bv \rangle + \frac{\tau}{2} \| b - Au_{k+1} - Bv \|^2 \\
    \lambda_{k+1} &= \lambda_k + \tau (b - Au_{k+1} - Bv_{k+1})
\end{align*}
\]
Penalty parameter

Constrained problem

\[
\min_{u,v} \ H(u) + G(v) \quad \text{subject to} \quad Au + Bv = b.
\]

Saddle point problem with augmented Lagrangian

\[
\max_\lambda \ \min_{u,v} \ H(u) + G(v) + \langle \lambda, b - Au - Bv \rangle + \frac{\tau}{2} \| b - Au - Bv \|^2
\]

How to select the free parameter in ADMM?

\[
u_{k+1} = \operatorname{arg\,min}_u \ H(u) + \langle \lambda_k, -Au \rangle + \frac{\tau_k}{2} \| b - Au - Bv_k \|^2
\]

\[
v_{k+1} = \operatorname{arg\,min}_v \ G(v) + \langle \lambda_k, -Bv \rangle + \frac{\tau_k}{2} \| b - Au_{k+1} - Bv \|^2
\]

\[
\lambda_{k+1} = \lambda_k + \tau_k \ (b - Au_{k+1} - Bv_{k+1})
\]
Background: spectral stepsize for gradient descent

Objective: \( \min_x F(x) \)

Gradient descent: \( x_{k+1} = x_k - \tau_k \nabla F(x_k) \)
Background: spectral stepsize for gradient descent

Objective: \(\min_x F(x)\)

Gradient descent: \(x_{k+1} = x_k - \tau_k \nabla F(x_k)\)

If quadratic, \(F(x) = \frac{\alpha}{2} \|x - x^*\|^2\), optimal \(\tau_k = 1/\alpha\)
Background: spectral stepsize for gradient descent

Objective: \( \min_x F(x) \)

Gradient descent: \( x_{k+1} = x_k - \tau_k \nabla F(x_k) \)

\[ x_k \]

\[ x_{k+1} \]

\[ \tau_k \]
Background: spectral stepsize for gradient descent

Gradient descent: \( x_{k+1} = x_k - \tau_k \nabla F(x_k) \)

Spectral (Barzilai-Borwen) stepsize: \( \tau_k = 1/\alpha \)
where \( \alpha \) is the local curvature assuming \( \nabla F(x) = \alpha x + a \)
Gradient descent: \( x_{k+1} = x_k - \tau_k \nabla F(x_k) \)

Spectral (Barzilai-Borwein) stepsize: \( \tau_k = 1/\alpha \) where \( \alpha \) is the local curvature assuming \( \nabla F(x) = \alpha x + a \) and \( \alpha \) is estimated by 1-dimensional least squares

\[
\nabla F(x_k) - \nabla F(x_{k-1}) = \alpha (x_k - x_{k-1})
\]

Background: spectral stepsize for gradient descent
Background: spectral stepsize for gradient descent

- Automates the stepsize selection
- Achieves fast convergence
- Constrained problem?

\[ \tau_k = 1 / \alpha \]
Dual interpretation of ADMM

Constrained problem

\[ \min_{u,v} H(u) + G(v) \quad \text{subject to} \quad Au + Bv = b. \]

Dual problem without constraints

\[ \min_{\lambda} H^*(A^T\lambda) - \langle \lambda, b \rangle \underbrace{+ G^*(B^T\lambda)}_{\hat{H}(\lambda)} \]

\[ \underbrace{+ G(\lambda)}_{\hat{G}(\lambda)} \]

\[ \lambda \]

---

\( F^* \) denotes the Fenchel conjugate of \( F \), defined as \( F^*(y) = \sup_x \langle x, y \rangle - F(x) \)
Dual interpretation of ADMM

Constrained problem

$$\min_{u,v} \ H(u) + G(v) \quad \text{subject to} \quad Au + Bv = b.$$ 

Dual problem \textbf{without constraints}

$$\min_{\lambda} \left[ \hat{H}(\lambda) + \hat{G}(\lambda) \right] = \min_{\lambda} \ H^*(A^T \lambda) - \langle \lambda, b \rangle + G^*(B^T \lambda),$$

Define \( \hat{\lambda}_{k+1} = \lambda_k + \tau_k (b - Au_{k+1} - Bv_k) \),

ADMM is equivalent to Douglas-Rachford Splitting (DRS) for dual

\( (u, v, \lambda) \Leftrightarrow (\hat{\lambda}, \lambda) \)

---

\( F^* \) denotes the Fenchel conjugate of \( F \), defined as \( F^*(y) = \sup_x \langle x, y \rangle - F(x) \).
Spectral stepsize of DRS

Dual problem \( \min_{\lambda} \left[ H^*(A^T \lambda) - \langle \lambda, b \rangle + G^*(B^T \lambda), \hat{H}(\lambda), \hat{G}(\lambda) \right] \)
Spectral stepsize of DRS

Dual problem \( \min_{\lambda} \underbrace{H^*(A^T \lambda) - \langle \lambda, b \rangle}_{\hat{H}(\lambda)} + \underbrace{G^*(B^T \lambda)}_{\hat{G}(\lambda)} \),

Approximate \( \partial \hat{H} \) and \( \partial \hat{G} \) at iteration \( k \) as linear functions

\[
\partial \hat{H}(\hat{\lambda}) = \alpha \hat{\lambda} + \Psi \quad \text{and} \quad \partial \hat{G}(\lambda) = \beta \lambda + \Phi
\]
Spectral stepsize of DRS

Dual problem

\[
\min_{\lambda} \quad \underbrace{H^*(A^T \lambda) - \langle \lambda, b \rangle}_{\hat{H}(\lambda)} + \underbrace{G^*(B^T \lambda)}_{\hat{G}(\lambda)},
\]

Approximate \( \partial \hat{H} \) and \( \partial \hat{G} \) at iteration \( k \) as linear functions

\[
\partial \hat{H}(\hat{\lambda}) = \alpha \hat{\lambda} + \Psi \quad \text{and} \quad \partial \hat{G}(\lambda) = \beta \lambda + \Phi
\]

[Proposition] When DRS is applied, the minimal residual of \( \hat{H}(\lambda_{k+1}) + \hat{G}(\lambda_{k+1}) \) is obtained by setting

\[
\tau_k = \frac{1}{\sqrt{\alpha \beta}}.
\]
**Spectral stepsize estimation**

- Spectral stepsize $\tau_k = 1/\sqrt{\alpha \beta}$
- Estimate curvature $\alpha, \beta$ of $\hat{H}, \hat{G}$ from ADMM iterates $(u, v, \hat{\lambda}, \lambda)$.
- 1-dimensional least squares with closed form solution.

$$A(u_k - u_{k0}) = \alpha \cdot (\hat{\lambda}_k - \hat{\lambda}_{k0})$$

$$B(v_k - v_{k0}) = \beta \cdot (\lambda_k - \lambda_{k0})$$

Recall $\hat{\lambda}_{k+1} = \lambda_k + \tau_k (b - Au_{k+1} - Bv_k)$.
Safeguarding inaccurate estimation

Objective: \( \min_x F(x) \)

Gradient descent: \( x_{k+1} = x_k - \tau_k \nabla F(x_k) \)

Backtracking linesearch
Safeguarding inaccurate estimation

Constrained problem

\[
\min_{u,v} \quad H(u) + G(v) \quad \text{subject to} \quad Au + Bv = b.
\]

Lagrangian saddle point problem

\[
\max_{\lambda} \min_{u,v} \quad H(u) + G(v) + \langle \lambda, b - Au - Bv \rangle
\]
Safeguarding

\[ A(u_k - u_{k_0}) (\hat{\lambda}_k - \hat{\lambda}_{k_0}) = \alpha \cdot \]

\[ B(v_k - v_{k_0}) (\lambda_k - \lambda_{k_0}) = \beta \cdot \]

\[ \alpha_{\text{cor}} = \frac{\langle A(u_k - u_{k_0}), \hat{\lambda}_k - \hat{\lambda}_{k_0} \rangle}{\|A(u_k - u_{k_0})\| \|\hat{\lambda}_k - \hat{\lambda}_{k_0}\|} \]

\[ \beta_{\text{cor}} = \frac{\langle B(v_k - v_{k_0}), \lambda_k - \lambda_{k_0} \rangle}{\|B(v_k - v_{k_0})\| \|\lambda_k - \lambda_{k_0}\|} \]

Validate correlations for the linear assumption of (sub)gradients.
Safeguarding

\( A(u_k - u_{k0}) \)

\( B(v_k - v_{k0}) \)

\( \alpha_{\text{cor}} \)

\( \beta_{\text{cor}} \)

Safeguarded spectral penalty parameter

\[
\tau_{k+1} = \begin{cases} 
\frac{1}{\sqrt{\alpha_k \beta_k}} & \text{if } \alpha_k^{\text{cor}} > \epsilon^{\text{cor}} \text{ and } \beta_k^{\text{cor}} > \epsilon^{\text{cor}} \\
\frac{1}{\alpha_k} & \text{if } \alpha_k^{\text{cor}} > \epsilon^{\text{cor}} \text{ and } \beta_k^{\text{cor}} \leq \epsilon^{\text{cor}} \\
\frac{1}{\beta_k} & \text{if } \alpha_k^{\text{cor}} \leq \epsilon^{\text{cor}} \text{ and } \beta_k^{\text{cor}} > \epsilon^{\text{cor}} \\
\tau_k & \text{otherwise},
\end{cases}
\]
Convergence guarantee

- Adaptive ADMM converges when one of the conditions is satisfied [He et al., 2000, Xu et al., 2017]
  - Bounded increasing
    \[
    \sum_{k=1}^{\infty} \eta_k^2 < \infty, \text{ where } \eta_k = \sqrt{\max\{\frac{\tau_k}{\tau_{k-1}}, 1\} - 1}
    \]
  - Bounded decreasing
    \[
    \sum_{k=1}^{\infty} \theta_k^2 < \infty, \text{ where } \theta_k = \sqrt{\max\{\frac{\tau_{k-1}}{\tau_k}, 1\} - 1}
    \]
Adaptive ADMM algorithm

- ADMM steps to update \((u_{k+1}, v_{k+1}, \lambda_{k+1})\)
- Estimate curvatures \(\alpha_k, \beta_k\)
- Estimate correlations \(\alpha^\text{cor}_k, \beta^\text{cor}_k\)
- Apply safeguarded spectral penalty rule to update \(\tau_{k+1}\)
- Stop adaptivity after fixed number of iterations to guarantee convergence
Experiments

- ADMM [Gabay and Mercier, 1976, Glowinski and Marroco, 1975, Boyd et al., 2011]
- Nesterov acceleration [Goldstein et al., 2014]
- Fixed optimal penalty parameter [Raghunathan and Di Cairano, 2014]
- Residual balancing [He et al., 2000, Boyd et al., 2011]

Residuals:

\[
\begin{align*}
    r_k &= b - Au_k - Bv_k, \\
    d_k &= \tau_k A^T B(v_k - v_{k-1})
\end{align*}
\]

\[
\tau_{k+1} = \begin{cases} 
    \eta \tau_k & \text{if } \|r_k\|_2 > \mu \|d_k\|_2 \\
    \tau_k / \eta & \text{if } \|d_k\|_2 > \mu \|r_k\|_2 \\
    \tau_k & \text{otherwise,}
\end{cases}
\]

\[
(\eta = 10, \mu = 2)
\]
Numerical results

- Applications: elastic net regularized linear regression; low rank least squares; dual SVM (quadratic programming); basis pursuit; consensus logistic regression; semidefinite programming

- Benchmark datasets from UCI repository and LIBSVM page.

- Initial penalty $\tau_0 = 0.1$, fixed safeguarding threshold $\epsilon_{\text{cor}} = 0.2$

- Details and more results in paper!

<table>
<thead>
<tr>
<th>Application</th>
<th>Dataset</th>
<th>Vanilla ADMM</th>
<th>Fast ADMM</th>
<th>Residual balance</th>
<th>Adaptive ADMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>Boston</td>
<td>2000+</td>
<td>208</td>
<td>54 (.023)</td>
<td>17 (.011)</td>
</tr>
<tr>
<td></td>
<td>Leukemia</td>
<td>2000+</td>
<td>2000+</td>
<td>1737 (19.3)</td>
<td>152 (1.70)</td>
</tr>
<tr>
<td>LRLS</td>
<td>Madelon</td>
<td>1943</td>
<td>193</td>
<td>133 (60.9)</td>
<td>27 (12.8)</td>
</tr>
<tr>
<td>Dual SVM</td>
<td>Madelon</td>
<td>100</td>
<td>57</td>
<td>28 (4.12)</td>
<td>19 (2.64)</td>
</tr>
<tr>
<td>BP</td>
<td>Human1</td>
<td>2000+</td>
<td>2000+</td>
<td>839 (.990)</td>
<td>503 (.626)</td>
</tr>
<tr>
<td>Consensus</td>
<td>Madelon</td>
<td>2000+</td>
<td>2000+</td>
<td>115 (42.1)</td>
<td>23 (20.8)</td>
</tr>
<tr>
<td></td>
<td>Realsim</td>
<td>1000+</td>
<td>1000+</td>
<td>121 (558)</td>
<td>22 (118)</td>
</tr>
<tr>
<td>SDP</td>
<td>Ham-11-2</td>
<td>2000+</td>
<td>2000+</td>
<td>1203 (4.15e3)</td>
<td>447 (1.49e3)</td>
</tr>
</tbody>
</table>
Residual plot

- Relative residual: \[
\text{max} \left\{ \frac{\|r_k\|_2}{\max\{\|Au_k\|_2,\|Bv_k\|_2,\|b\|_2\}}, \frac{\|d_k\|_2}{\|A^T \lambda_k\|_2} \right\}
\]

- Low rank least squares: \[
\min_X \frac{1}{2} \|DX - C\|_F^2 + \rho_1 \|X\|_* + \frac{\rho_2}{2} \|X\|_F^2
\]

- Vanilla ADMM, Fast ADMM, Residual balance, Adaptive ADMM
Sensitivity: initial penalty

- Elastic net regularized linear regression

\[
\min_{u,v} \frac{1}{2} \|Du - c\|_2^2 + \rho_1 \|v\|_1 + \frac{\rho_2}{2} \|v\|_2^2 \quad \text{s.t.} \quad u - v = 0
\]

- Vanilla ADMM, Fast ADMM, Residual balance, Adaptive ADMM

![Graph showing iterations vs. initial penalty parameter for different methods]
Sensitivity: problem scale

- Quadratic programming

\[
\min_{u,v} \frac{1}{2} u^T s^2 Qu + s q^T u + \nu_{\{z: z_i \leq c\}}(v) \quad \text{s.t.} \quad Du - v = 0
\]

- Vanilla ADMM, Fast ADMM, Residual balance, Adaptive ADMM

![Graph showing iterations vs. problem scale for different algorithms](image)
Sensitivity: safeguarding threshold

- $\epsilon^{\text{cor}} = 0.2$ works well
Conclusion and extensions

Spectral penalty parameter selection for constrained problem
ADMM is equivalent to DRS of unconstrained dual problem
Combine the estimated curvatures of the two functions
Effective safeguarding
Fully automated and fast convergence

Relaxed ADMM [Xu et al., 2017]
Nonconvex applications [Xu et al., 2016b]
Multi-block ADMM [Xu et al., under review]
Large-scale distributed computing [Xu et al., under review]
O(1/k) convergence rate [Xu et al., under review]
Thank you!


