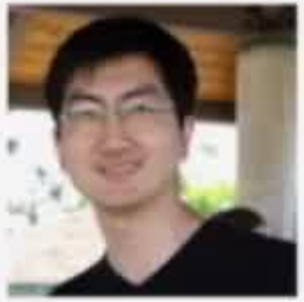
Adaptive ADMM with Spectral Penalty Parameter Selection

Zheng Xu

xuzhustc@gmail.com



T. Goldstein M. Figueiredo X. Yuan H. Li G. Taylor C. Studer

UNIVERSITY OF MARYLAND 18 56

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Constrained problem

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

Constrained problem

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

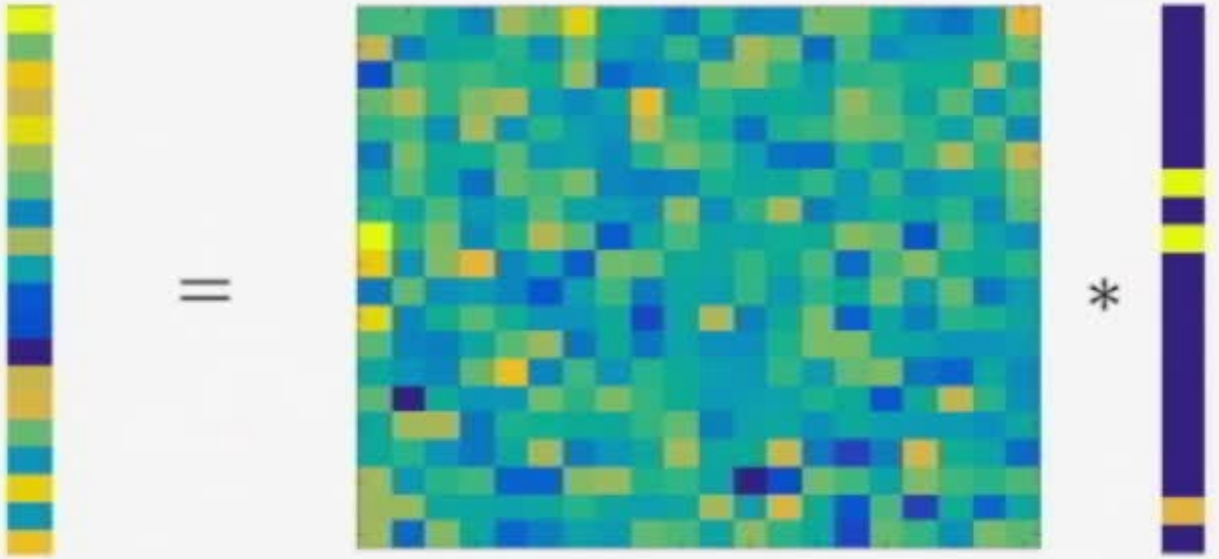
- Quadratic programming
- Semidefinite programming
- Sparse linear regression (L1 norm)
- Basis pursuit
- Low rank problem (nuclear norm)
- Robust principal component analysis (RPCA)
- Support vector machine (SVM)
- Total variation image denoising
- Phase retrieval
- Distributed computing …

Statistical learning problem

$$\min_v f(v) + g(v)$$

- Example:

$$\min_x \frac{1}{2}\|Dx - c\|_2^2 + \rho_1\|x\|_1 + \frac{\rho_2}{2}\|x\|_2^2$$

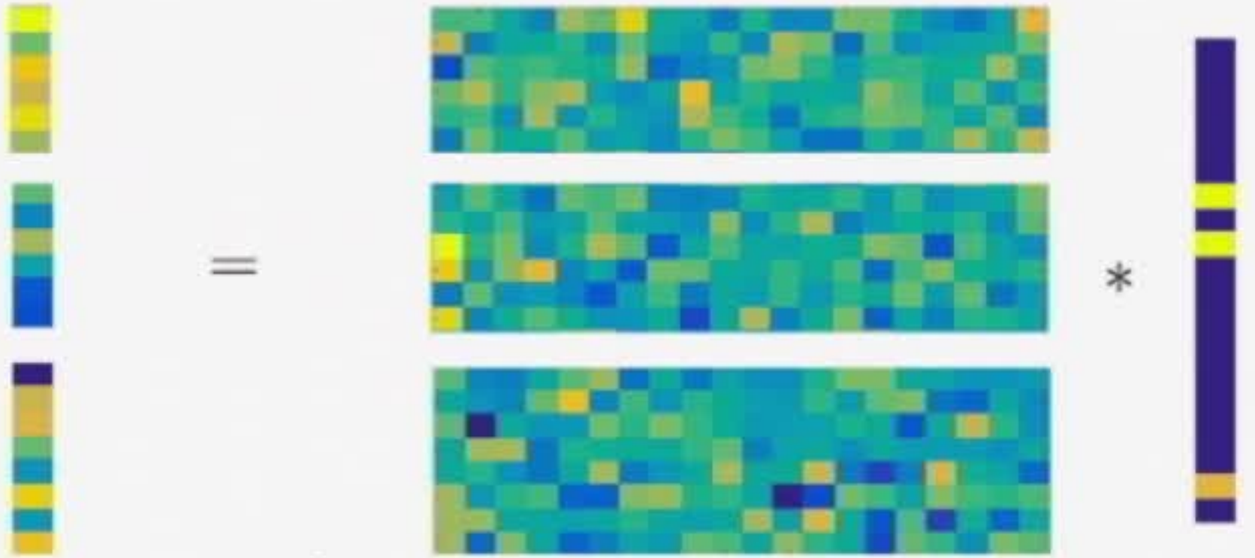


Problem decomposition and data parallelism

$$\min_v \sum_{i=1}^{N} f_i(v) + g(v)$$

- Example:

$$\min \sum_{i=1}^{N} \frac{1}{2}\|D_i x - c_i\|^2 + \rho_1 |x| + \frac{\rho_2}{2}\|x\|^2$$



$c = [c_1; \ldots; c_i; \ldots; c_N]$ $D = [D_1; \ldots; D_i; \ldots; D_N]$ x

Consensus problem

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

$$\min_{u_i,v} \; \sum_{i=1}^{N} f_i(u_i) + g(v), \quad \text{subject to} \quad u_i = v$$

- Example:

local nodes

$$f_i(u_i) = \frac{1}{2}\|D_i u_i - c_i\|^2$$

central server

$$g(v) = \rho_1 |v| + \frac{\rho_2}{2}\|v\|^2$$

How to solve constrained problem?

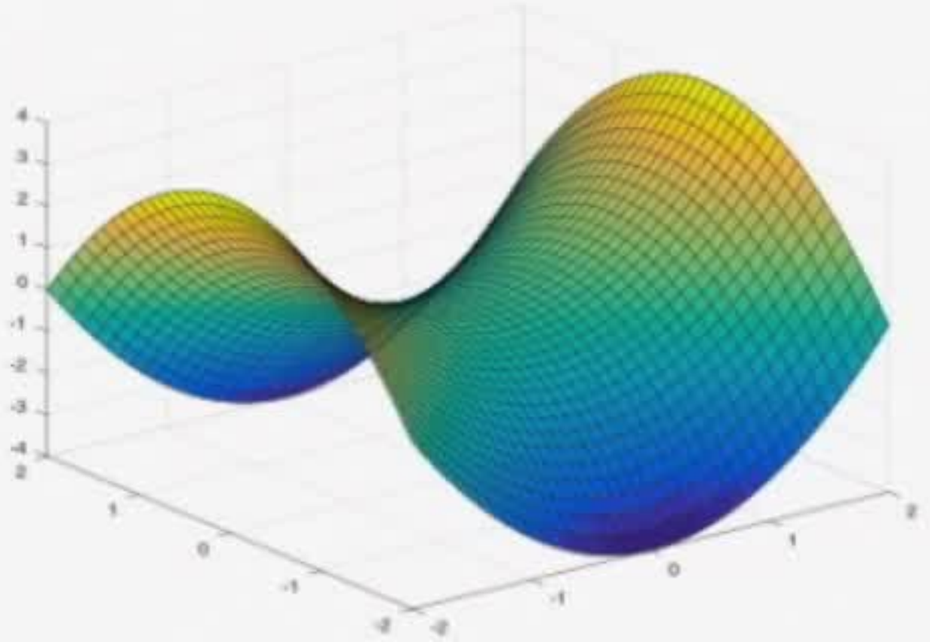
$$\min_{u,v} \; H(u) + G(v) \text{ subject to } Au + Bv = b$$

How to solve constrained problem?

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

- Saddle point problem with augmented Lagrangian

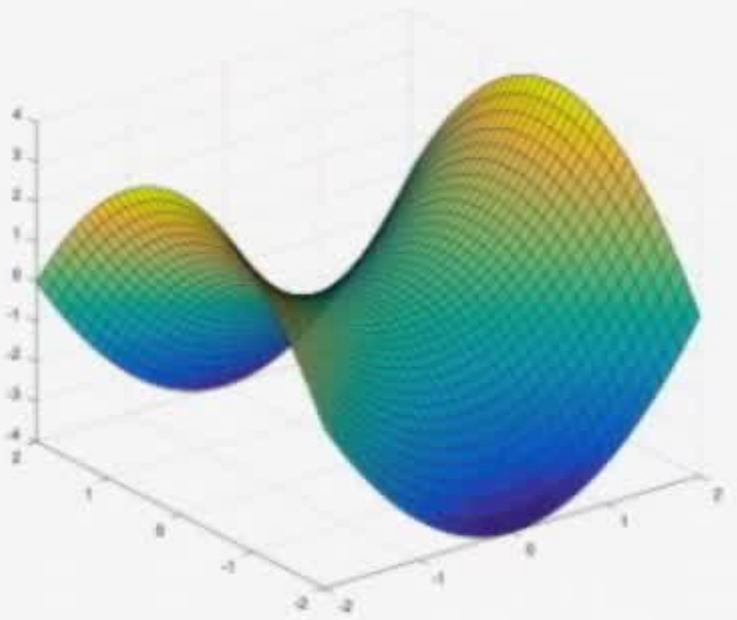
$$\max_{\lambda} \min_{u,v} H(u) + G(v) + \langle \lambda, b - Au - Bv \rangle + \frac{\tau}{2} \|b - Au - Bv\|^2$$



Penalty parameter

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

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



$$u_{k+1} = \arg\min_{u} H(u) + \langle \lambda_k, -Au \rangle + \frac{\tau_k}{2} \|b - Au - Bv_k\|^2$$

$$v_{k+1} = \arg\min_{v} G(v) + \langle \lambda_k, -Bv \rangle + \frac{\tau_k}{2} \|b - Au_{k+1} - Bv\|^2$$

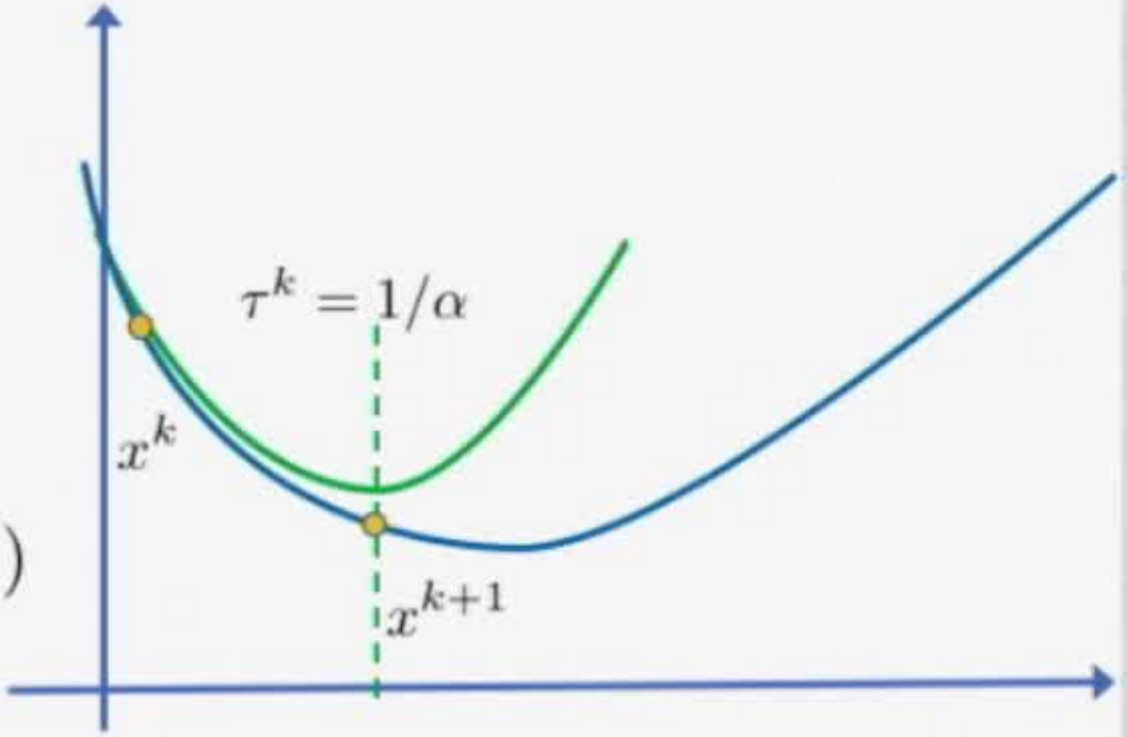
$$\lambda_{k+1} = \lambda_k + \tau_k (b - Au_{k+1} - Bv_{k+1})$$

The only free parameter!

Background: spectral stepsize

- Objective $\min_x F(x)$
- Gradient descent $x^{k+1} = x^k - \tau^k \nabla F(x^k)$

- Spectral (Barzilai-Borwein) stepsize
  - Assume the function is locally quadratic with curvature α
  - (Sub)gradient is linear $\nabla F(x) = \alpha x + a$
  - Estimate the curvature by solving 1-d least squares
    $\nabla F(x^k) - \nabla F(x^{k-1}) = \alpha(x^k - x^{k-1})$
  - Gradient descent with $\tau^k = 1/\alpha$

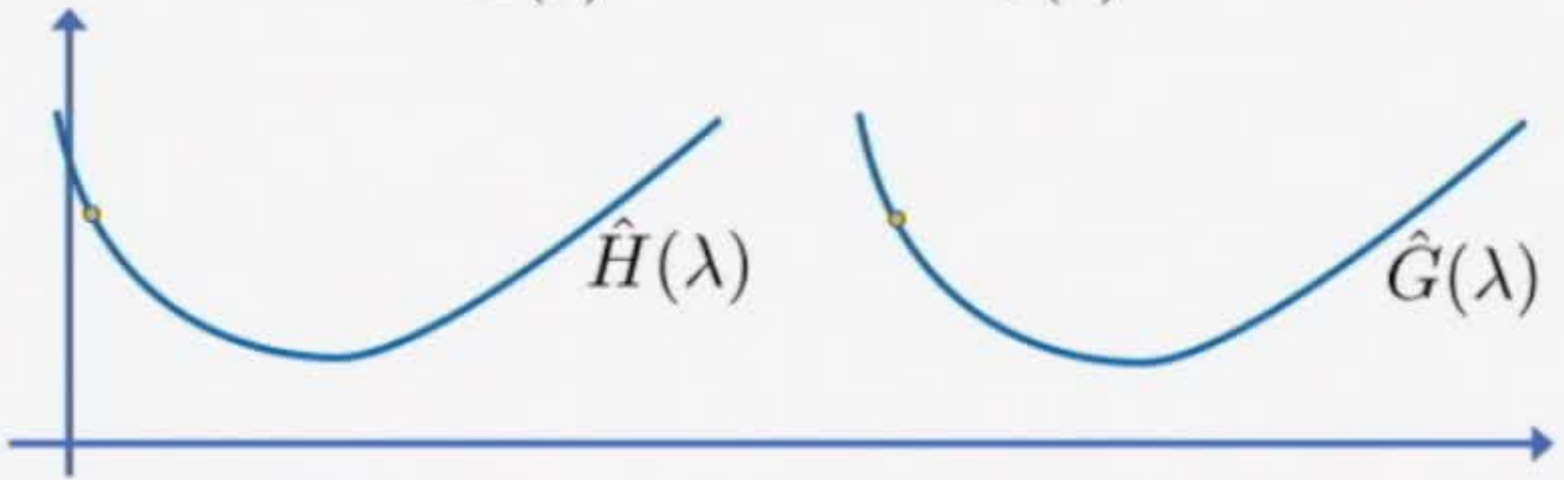


J. Barzilai and J. Borwein. Two-point step size gradient methods. 1988.

Dual interpretation

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

- Dual problem $\min_{\lambda} \underbrace{H^*(A^T\lambda) - \langle \lambda, b \rangle}_{\hat{H}(\lambda)} + \underbrace{G^*(B^T\lambda)}_{\hat{G}(\lambda)}$ No constraints!



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

Dual problem and DRS

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

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

ADMM (u, v, λ) $\Longleftrightarrow$ Douglas-Rachford Splitting $(\hat{\lambda}, \lambda)$

Defininig

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

$$u_{k+1} = \arg\min_u H(u) + \langle \lambda_k, -Au\rangle + \frac{\tau_k}{2}\|b - Au - Bv_k\|^2$$

$$v_{k+1} = \arg\min_v G(v) + \langle \lambda_k, -Bv\rangle + \frac{\tau_k}{2}\|b - Au_{k+1} - Bv\|^2$$

$$\lambda_{k+1} = \lambda_k + \tau_k(b - Au_{k+1} - Bv_{k+1})$$

$$0 \in \frac{\hat{\lambda}_{k+1} - \lambda_k}{\tau_k} + \partial\hat{H}(\hat{\lambda}_{k+1}) + \partial\hat{G}(\lambda_k)$$

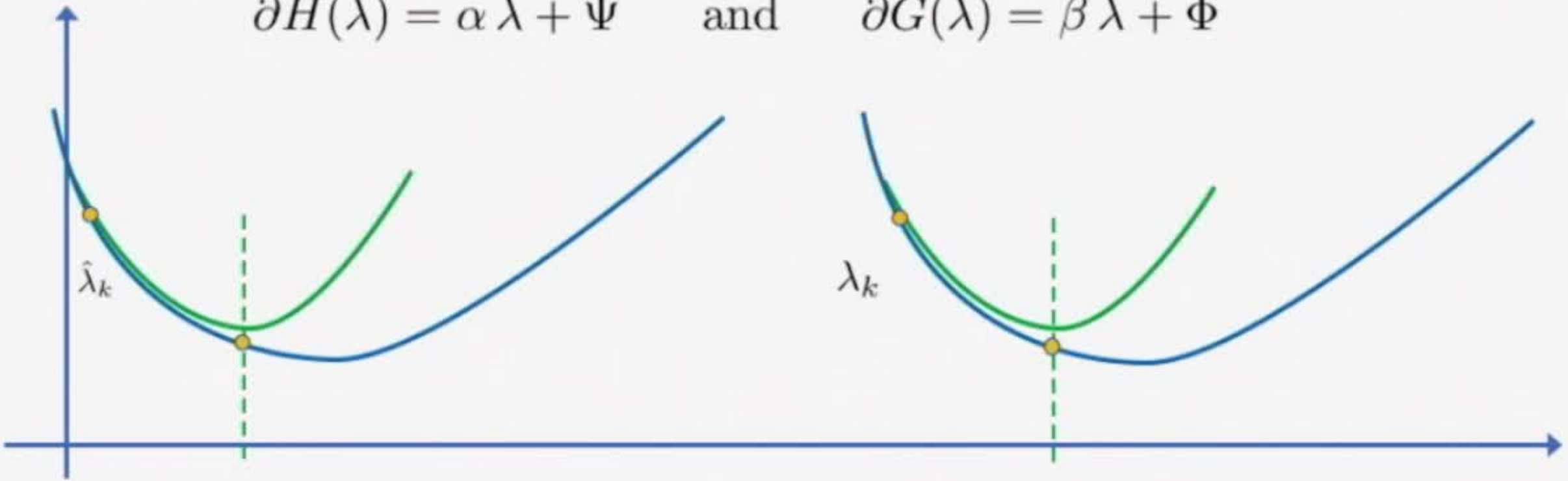
$$0 \in \frac{\lambda_{k+1} - \zeta_k}{\tau_k} + \partial\hat{H}(\hat{\lambda}_{k+1}) + \partial\hat{G}(\lambda_{k+1})$$

Linear approximation

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

- Approximate subgradients of dual functions at iteration k as linear functions

$$\partial \hat{H}(\hat{\lambda}) = \alpha\,\hat{\lambda} + \Psi \qquad \text{and} \qquad \partial \hat{G}(\lambda) = \beta\,\lambda + \Phi$$



Spectral stepsize

Proposition (Spectral DRS/ADMM).
Suppose the DRS steps are applied to dual problem,

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

where (omitting the subscript k from $\alpha_k, \beta_k, \Psi_k, \Phi_k$ to lighten the notation)

$$\partial\hat{H}(\hat{\lambda}) = \alpha\,\hat{\lambda} + \Psi \qquad \textit{and} \qquad \partial\hat{G}(\lambda) = \beta\,\lambda + \Phi.$$

Then, the minimal residual of $\hat{H}(\lambda_{k+1}) + \hat{G}(\lambda_{k+1})$ is obtained by setting

$$\tau_k = 1/\sqrt{\alpha\,\beta}\,.$$

Z. Xu, M. Figueiredo, and T. Goldstein. Adaptive ADMM with spectral penalty parameter selection. AISTATS, 2017.

Estimating spectral penalty parameter

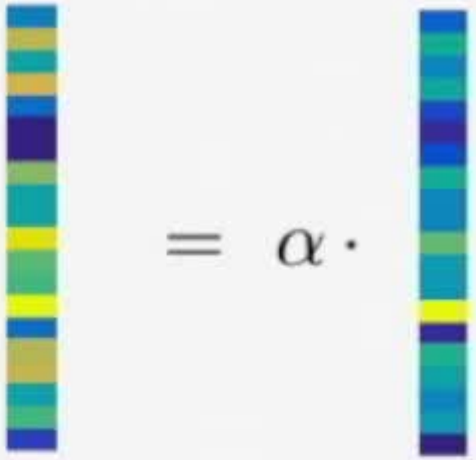
- Spectral penalty schema

$$\tau_k = 1/\sqrt{\alpha\,\beta}$$

- Estimate curvatures $\alpha,\ \beta$ from ADMM iterates $(u, v, \hat{\lambda}, \lambda)$
  1-dimensional least squares with closed form solution

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

$\partial\hat{H}(\hat{\lambda}_k) - \partial\hat{H}(\hat{\lambda}_{k_0})$



$= \alpha \cdot$

$\partial\hat{G}(\lambda_k) - \partial\hat{G}(\lambda_{k_0})$

$= \beta \cdot$

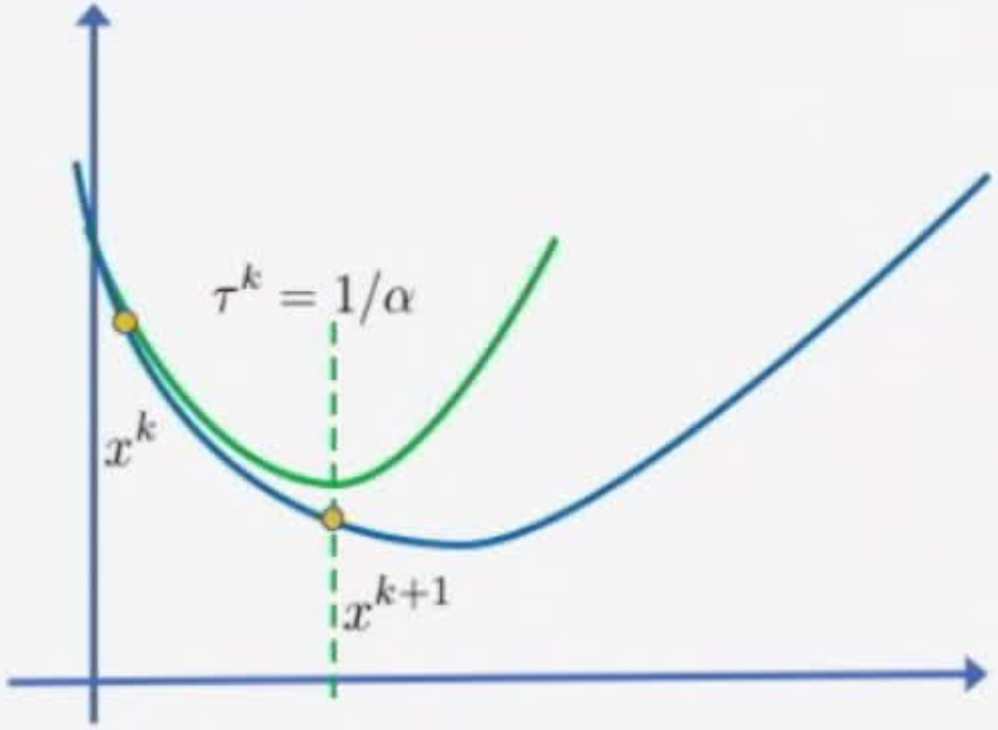
$A(u_k - u_{k_0})$ $(\hat{\lambda}_k - \hat{\lambda}_{k_0})$ $B(v_k - v_{k_0})$ $(\lambda_k - \lambda_{k_0})$

Safeguarding inaccurate estimation

- Objective $\min_x F(x)$
- Gradient descent

$$x^{k+1} = x^k - \tau^k \nabla F(x^k)$$

- Line search

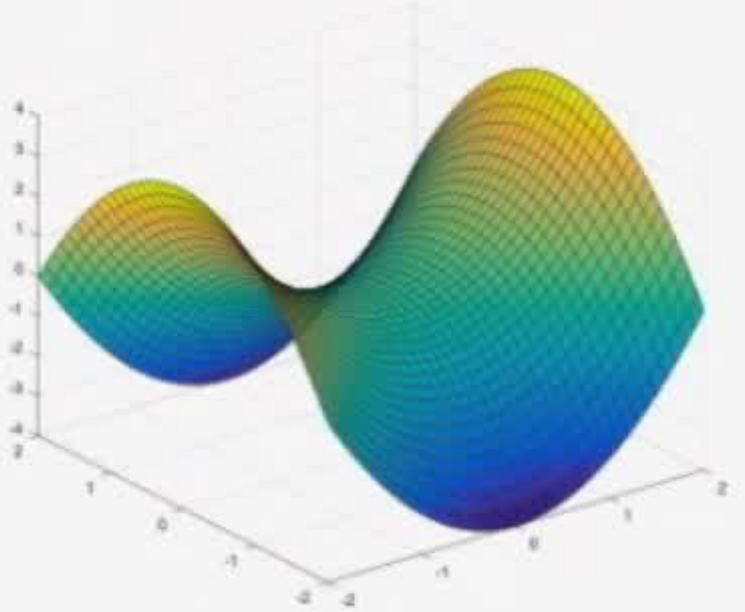


- Constrained problem

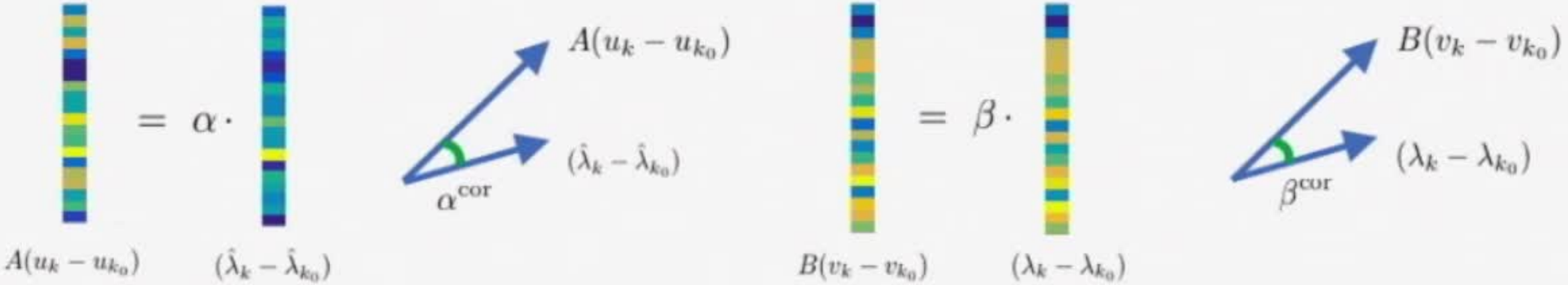
$$\min_{u,v} \; H(u) + G(v)$$

$$\text{subject to } Au + Bv = b$$

- ADMM
- How to safeguard?



Safeguarding by correlation



- Safeguarded spectral penalty schema

$$\tau_{k+1} = \begin{cases} 1/\sqrt{\alpha_k \beta_k} & \text{if } \alpha_k^{\mathrm{cor}} > \epsilon^{\mathrm{cor}} \text{ and } \beta_k^{\mathrm{cor}} > \epsilon^{\mathrm{cor}} \\ 1/\alpha_k & \text{if } \alpha_k^{\mathrm{cor}} > \epsilon^{\mathrm{cor}} \text{ and } \beta_k^{\mathrm{cor}} \leq \epsilon^{\mathrm{cor}} \\ 1/\beta_k & \text{if } \alpha_k^{\mathrm{cor}} \leq \epsilon^{\mathrm{cor}} \text{ and } \beta_k^{\mathrm{cor}} > \epsilon^{\mathrm{cor}} \\ \tau_k & \text{otherwise,} \end{cases}$$

O (1/k) convergence with adaptivity

- Bounded adaptivity $\sum_{k=1}^{\infty} \eta_k^2 < \infty, \text{ where } \eta_k^2 = \max_{i \in \{1,\dots,p\}} \{\eta_{i,k}^2\},$

$$\eta_{i,k}^2 = \max\{\tau_{i,k}/\tau_{i,k-1} - 1, \tau_{i,k-1}/\tau_{i,k} - 1\}.$$

- The norm of the residuals converges to zero
- The worst-case ergodic O(1/k) convergence rate in the variational inequality sense

Z. Xu, G. Taylor, H. Li, M. Figueiredo, X. Yuan, and T. Goldstein. Adaptive consensus ADMM for distributed optimization. ICML, 2017

Experiments

- ADMM methods
  - Vanilla ADMM
  - Fixed optimal penalty [Raghunathan 2014]
  - Residual balancing [He et al. 2000]
  - Fast ADMM [Goldstein et al. 2014]
  - Adaptive ADMM (AADMM)

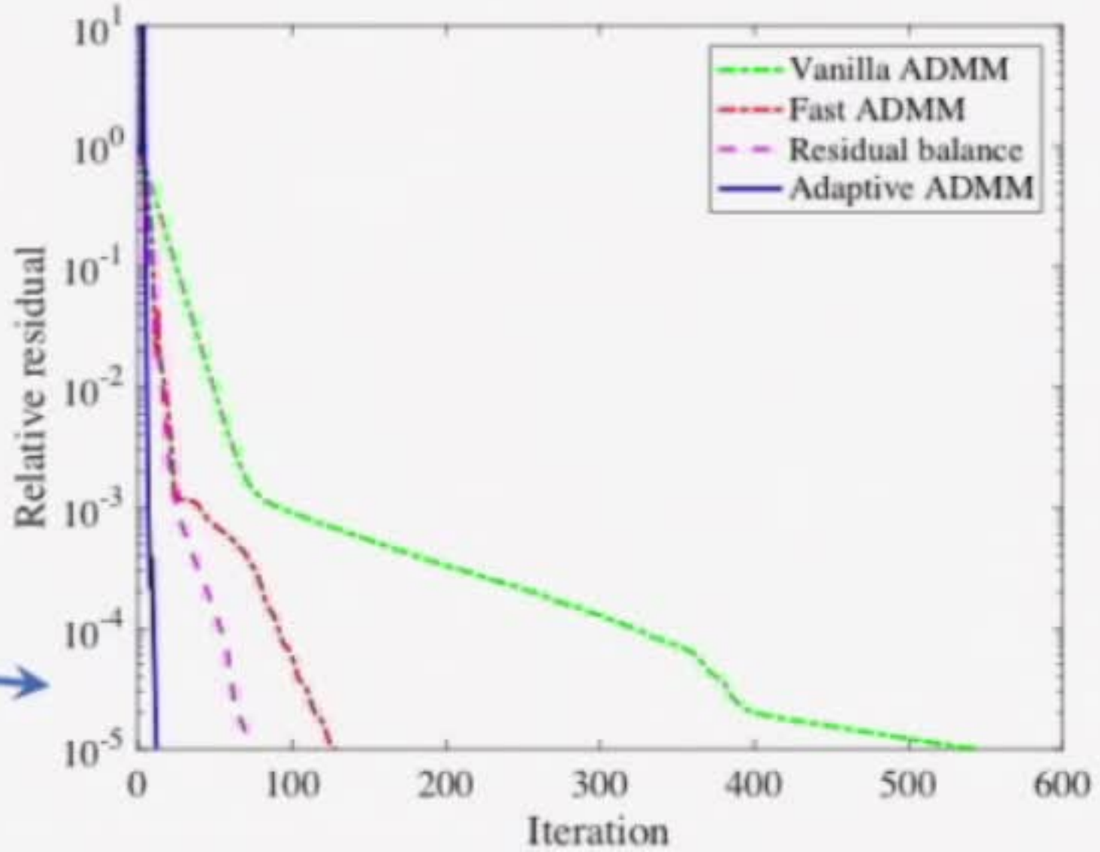
- Applications
  - Linear regression with elastic net regularizer
  - Low rank least squares
  - Support vector machine
  - Basis pursuit
  - Consensus logistic regression
  - Semidefinite programming

Residual plot

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

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

AADMM →



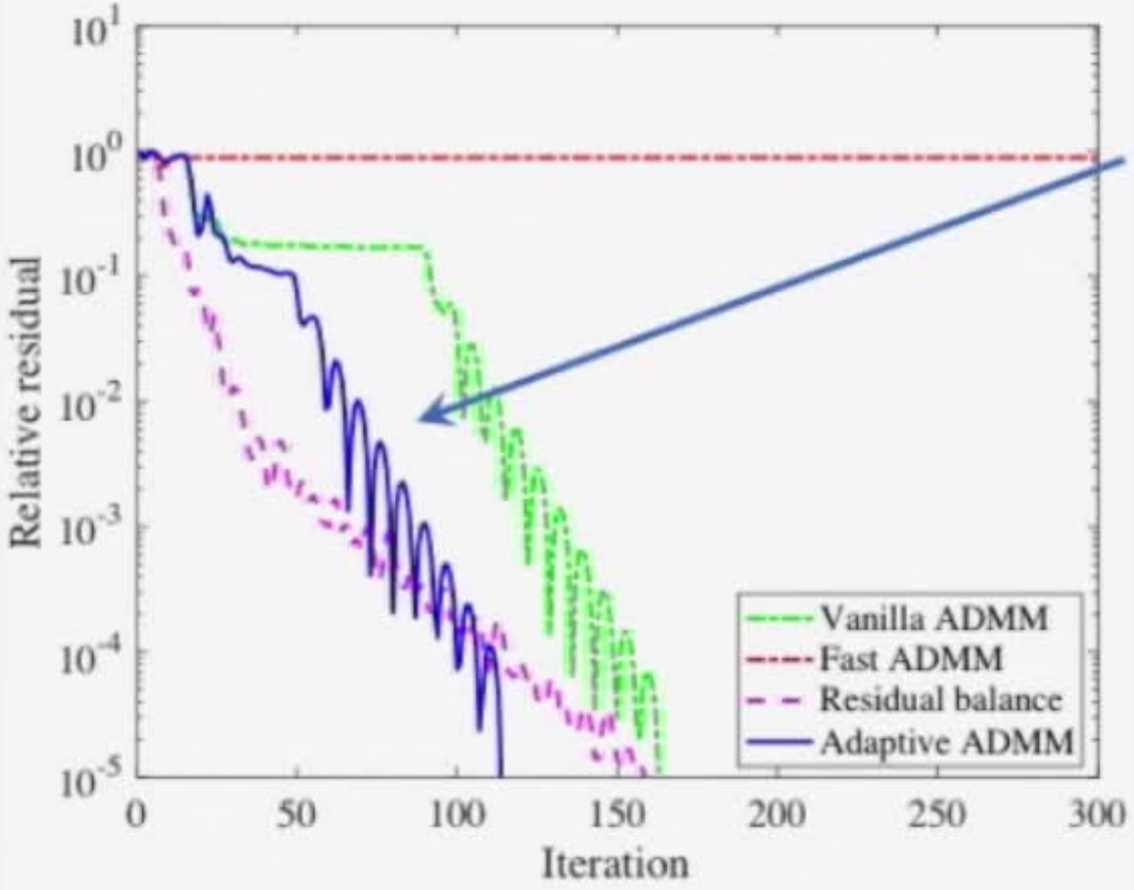
Experiments

- ADMM methods
  - Vanilla ADMM
  - Fixed optimal penalty [Raghunathan 2014]
  - Residual balancing [He et al. 2000]
  - Fast ADMM [Goldstein et al. 2014]
  - Adaptive ADMM (AADMM)

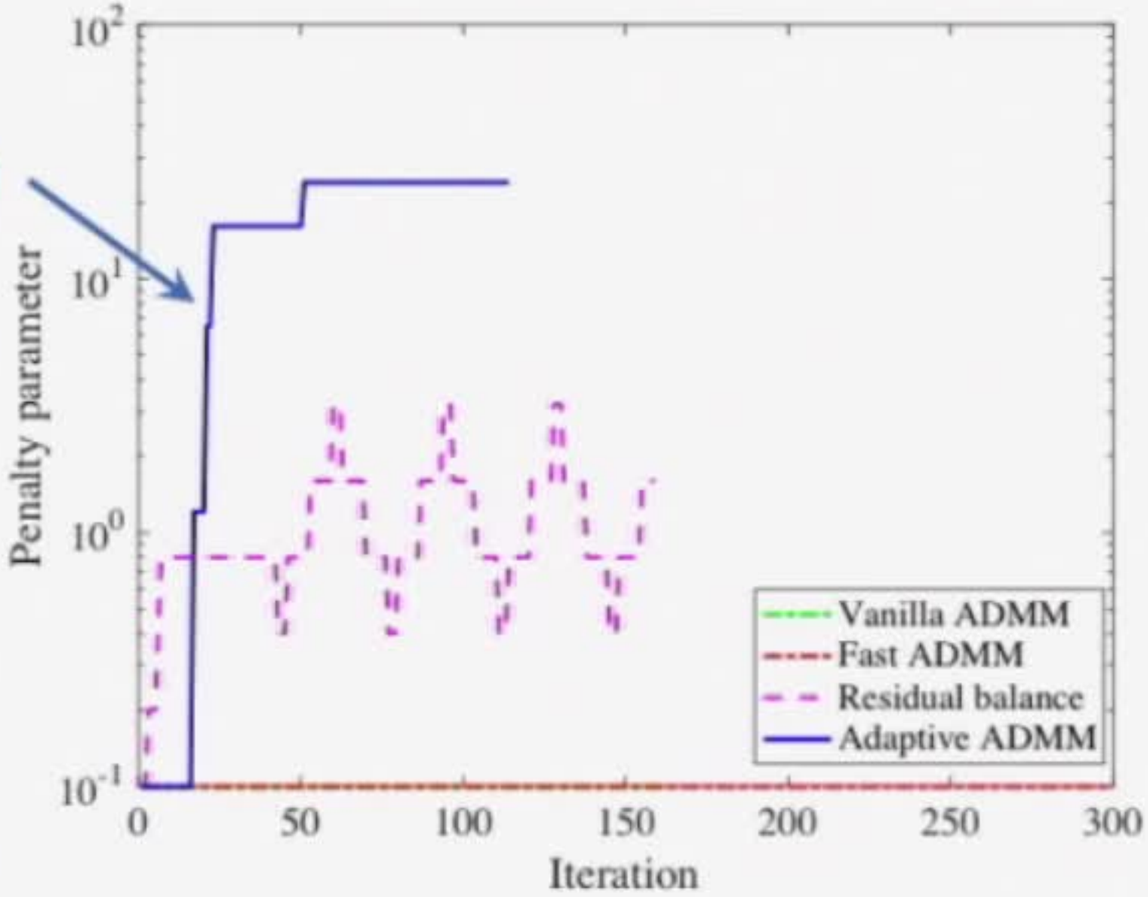
- Applications
  - Linear regression with elastic net regularizer
  - Low rank least squares
  - Support vector machine
  - Basis pursuit
  - Consensus logistic regression
  - Semidefinite programming

Residual plot II

- Basis pursuit $\min_x \|x\|_1 \quad \text{subject to} \quad Dx = c,$



AADMM



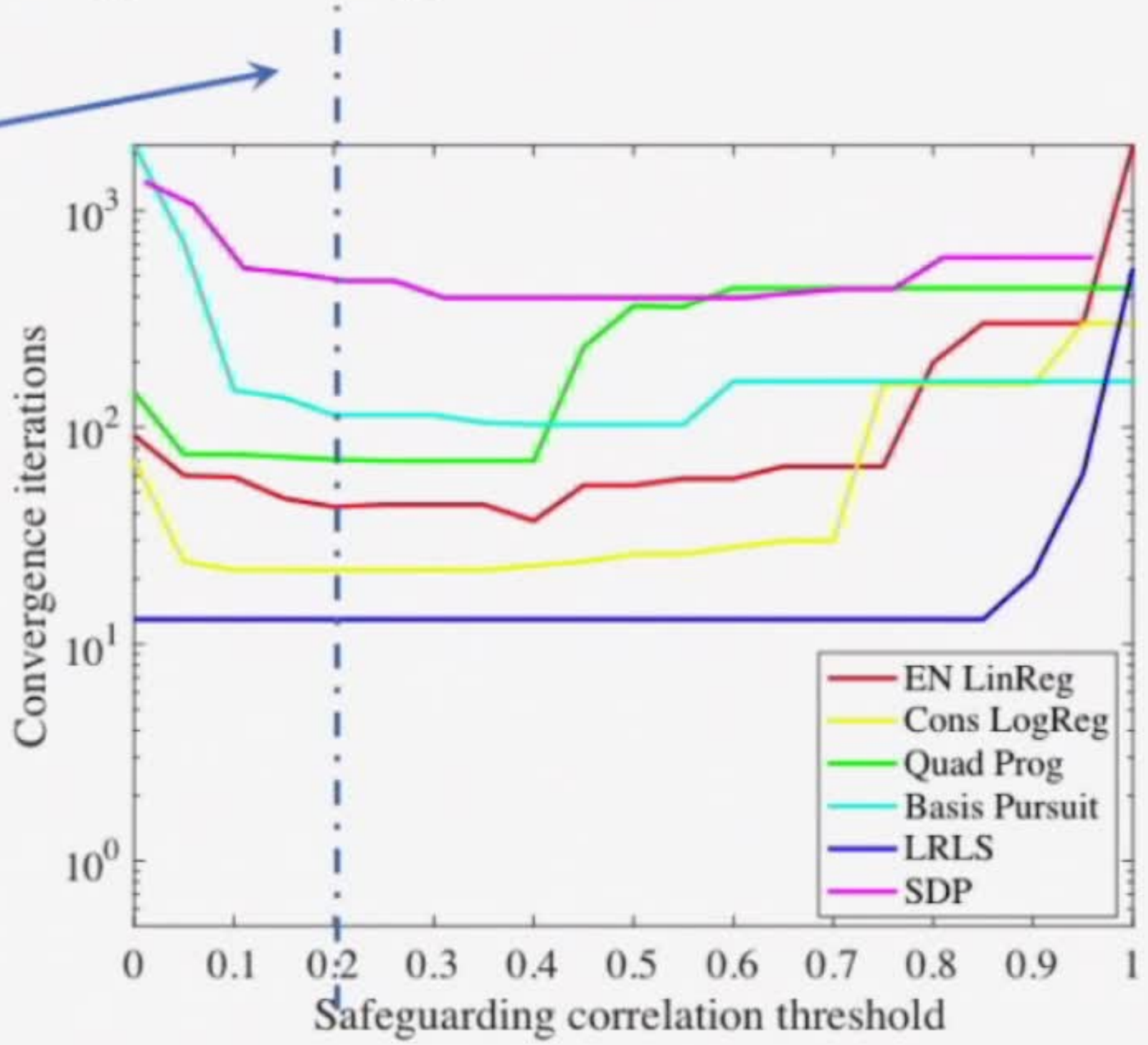
Convergence results

- Benchmark datasets from UCI repository and LIBSVM page.
- Number of iterations (runtime in seconds)

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

Robust to safeguarding threshold

0.2 is used for all other experiments



More experiments

Application	Dataset	#samples × #features[3]	Vanilla ADMM	Fast ADMM	Residual balance	**Adaptive ADMM**
Elastic net regression	Synthetic	50 × 40	2000+ (1.64)	263 (.270)	111 (.129)	**43 (.046)**
	Boston	506 × 13	2000+ (2.19)	208 (.106)	54 (.023)	**17 (.011)**
	Diabetes	768 × 8	594 (.269)	947 (.848)	28 (.020)	**10 (.005)**
	Leukemia	38 × 7129	2000+ (22.9)	2000+ (24.2)	1737 (19.3)	**152 (1.70)**
	Prostate	97 × 8	548 (.293)	139 (.049)	29 (.015)	**16 (.012)**
	Servo	130 × 4	142 (.040)	44 (.017)	27 (.012)	**13 (.007)**
Low rank least squares	Synthetic	1000 × 200	543(31.3)	129(7.30)	75(5.59)	**13(.775)**
	Madelon	2000 × 500	1943(925)	193(89.6)	133(60.9)	**27(12.8)**
	Sonar	208 × 60	1933(9.12)	313(1.51)	102(.466)	**31(.160)**
	Splice	1000 × 60	1704(38.2)	189(4.25)	92(2.04)	**18(.413)**
QP[3] and dual SVM	Synthetic	250 × 500	439 (6.15)	535 (7.8380)	232 (3.27)	**71 (.984)**
	Madelon	2000 × 500	100 (14.0)	57 (8.14)	28 (4.12)	**19 (2.64)**
	Sonar	208 × 60	139 (.227)	43 (.075)	37 (.069)	**28 (.050)**
	Splice	1000 × 60	149 (4.9)	47 (1.44)	39 (1.27)	**20 (.681)**
Basis pursuit	Synthetic	10 × 30	163 (.027)	2000+ (.310)	159 (.031)	**114 (.026)**
	Human1	1024 × 1087	2000+ (2.35)	2000+ (2.41)	839 (.990)	**503 (.626)**
	Human2	1024 × 1087	2000+ (2.26)	2000+ (2.42)	875 (1.03)	**448 (.554)**
	Human3	1024 × 1087	2000+ (2.29)	2000+ (2.44)	713 (.855)	**523 (.641)**
Consensus logistic regression	Synthetic	1000 × 25	301 (3.36)	444 (3.54)	43 (.583)	**22 (.282)**
	Madelon	2000 × 500	2000+ (205)	2000+ (166)	115 (42.1)	**23 (20.8)**
	Sonar	208 × 60	2000+ (33.5)	2000+ (47)	106 (2.82)	**90 (1.64)**
	Splice	1000 × 60	2000+ (29.1)	2000+ (43.7)	86 (1.91)	**22 (.638)**
	News20	19996 × 1355191	69 (5.91e3)	32 (3.45e3)	18 (1.52e3)	**16 (1.2e3)**
	Rcv1	20242 × 47236	38 (177)	23 (122)	**13 (53.0)**	**12 (53.9)**
	Realsim	72309 × 20958	1000+ (2.73e3)	1000+ (1.86e3)	121 (558)	**22 (118)**
Semidefinite programming	hamming-7-5-6	128 × 1792	455(1.78)	2000+(8.60)	1093(4.21)	**284(1.11)**
	hamming-8-3-4	256 × 16128	418(6.38)	2000+(29.1)	1071(16.5)	**118(2.02)**
	hamming-9-5-6	512 × 53760	2000+(187)	2000+(187)	1444(131)	**481(53.1)**
	hamming-9-8	512 × 2304	2000+(162)	2000+(159)	1247(97.2)	**594(52.7)**
	hamming-10-2	1024 × 23040	2000+(936)	2000+(924)	1194(556)	**391(193)**
	hamming-11-2	2048 × 56320	2000+(6.43e3)	2000+(6.30e3)	1203(4.15e3)	**447(1.49e3)**

[3] #constrains × #unknowns for canonical QP; #vertices × #edges for SDP.

Application	Dataset	#samples × #features[3]	Vanilla ADMM	Relaxed ADMM	Residual balance	Adaptive ADMM	**Proposed ARADMM**
Elastic net regression	Synthetic	50 × 40	2000+(.642)	2000+(.660)	424(.144)	102(.051)	**70(.026)**
	MNIST	60000 × 784	1225(29.4)	816(19.9)	94(2.28)	41(.943)	**21(.549)**
	CIFAR10	10000 × 3072	2000+(690)	2000+(697)	556(193)	2000+(669)	**94(31.7)**
	News20	19996 × 1355191	2000+(1.21e4)	2000+(9.16e3)	227(914)	104(391)	**71(287)**
	Rcv1	20242 × 47236	2000+(1.20e3)	1823(802)	196(79.1)	104(35.7)	**64(26.0)**
	Realsim	72309 × 20958	2000+(4.26e3)	2000+(4.33e3)	341(355)	152(125)	**107(88.2)**
Low rank least squares	Synthetic	1000 × 200	2000+(118)	2000+(116)	268(15.1)	26(1.55)	**18(1.04)**
	German	1000 × 24	2000+(4.72)	2000+(4.72)	642(1.52)	130(.334)	**52(.125)**
	Spectf	80 × 44	2000+(2.70)	2000+(2.74)	336(.455)	162(.236)	**105(.150)**
	MNIST	60000 × 784	200+(1.86e3)	200+(2.08e3)	200+(3.29e3)	200+(3.46e3)	**38(658)**
	CIFAR10	10000 × 3072	200+(7.24e3)	200+(1.33e4)	53(1.60e3)	8(208)	**6(156)**
QP and dual SVM	Synthetic	250 × 500	1224(11.5)	823(7.49)	626(5.93)	170(1.57)	**100(.914)**
	German	1000 × 24	2000+(58.8)	2000+(61.8)	1592(45.0)	1393(38.9)	**1238(34.9)**
	Spectf	80 × 44	2000+(.846)	2000+(.777)	169(.070)	175(.086)	**53(.026)**
Consensus logistic regression	Synthetic	1000 × 25	590(9.93)	391(6.97)	70(1.23)	35(.609)	**20(.355)**
	German	1000 × 24	2000+(34.3)	2000+(66.6)	151(2.60)	35(.691)	**26(.580)**
	Spectf	80 × 44	1005(20.1)	667(14.4)	117(1.98)	145(1.63)	**85(1.07)**
	MNIST	60000 × 784	200+(2.99e3)	200+(3.47e3)	200+(1.37e3)	49(536)	**28(333)**
	CIFAR10	10000 × 3072	200+(593)	200+(2.08e3)	200+(1.54e3)	131(165)	**19(33.7)**
Unwrapping SVM	Synthetic	1000 × 25	2000+(1.13)	1418(.844)	2000+(1.16)	355(.229)	**147(.094)**
	German	1000 × 24	753(1.88)	560(1.37)	2000+(4.98)	572(1.44)	**213(.545)**
	Spectf	80 × 44	567(.203)	367(.112)	567(.185)	207(.068)	**149(.052)**
	MNIST	60000 × 784	128(130)	118(111)	163(153)	200+(217)	**67(71.0)**
	CIFAR10	10000 × 3072	200+(512)	200+(532)	200+(516)	89(285)	**57(143)**
Image denoising	Barbara	512 × 512	262(35.0)	175(23.6)	74(10.0)	59(8.67)	**38(5.57)**
	Cameraman	256 × 256	311(8.96)	208(5.89)	82(2.29)	88(2.76)	**35(1.08)**
	Lena	512 × 512	347(46.3)	232(31.3)	94(12.5)	68(9.70)	**39(5.58)**
Robust PCA	FaceSet1	64 × 1024	2000+(41.1)	1507(30.3)	560(11.1)	561(11.9)	**267(5.65)**
	FaceSet2	64 × 1024	2000+(41.1)	2000+(41.4)	263(5.54)	388(9.00)	**188(4.02)**
	FaceSet3	64 × 1024	2000+(39.4)	1843(36.3)	375(7.44)	473(9.89)	**299(6.27)**

[3] #constrains × #unknowns for canonical QP; width × height for image restoration.

Z. Xu, M. Figueiredo, and T. Goldstein. Adaptive ADMM with spectral penalty parameter selection. AISTATS, 2017.

Z. Xu, S. De, M. Figueiredo, C. Studer and T. Goldstein. An empirical study of ADMM for nonconvex problems. NIPS workshop, 2016.

Z. Xu, M. Figueiredo, X. Yuan, C. Studer, and T. Goldstein. Adaptive relaxed ADMM: convergence theory and practical implementation. CVPR, 2017.

Z. Xu, G. Taylor, H. Li, M. Figueiredo, X. Yuan, and T. Goldstein. Adaptive consensus ADMM for distributed optimization. ICML, 2017

Summary

- Spectral penalty parameter for constrained problem
- ADMM is equivalent to DRS of unconstrained dual problem
- Simple schema to combine the curvature estimations
- Effective safeguarding
- O(1/k) convergence rate
- Fully automated and fast practical convergence
- Implementation for more than ten different applications
- Code https://sites.google.com/site/xuzhustc/