LAGRANGIAN METHODS

WHY LAGRANGIAN METHODS?

Smooth functions

$$\text{minimize} \quad f(x)$$

→ Gradient descent
Newton's method
Quasi-newton
Conjugate gradients
etc…

Non-differentiable

$$\text{minimize} \quad f(x)$$

→ Proximal methods

Constrained problems?

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

→ **Lagrangian methods**

LAGRANGIAN

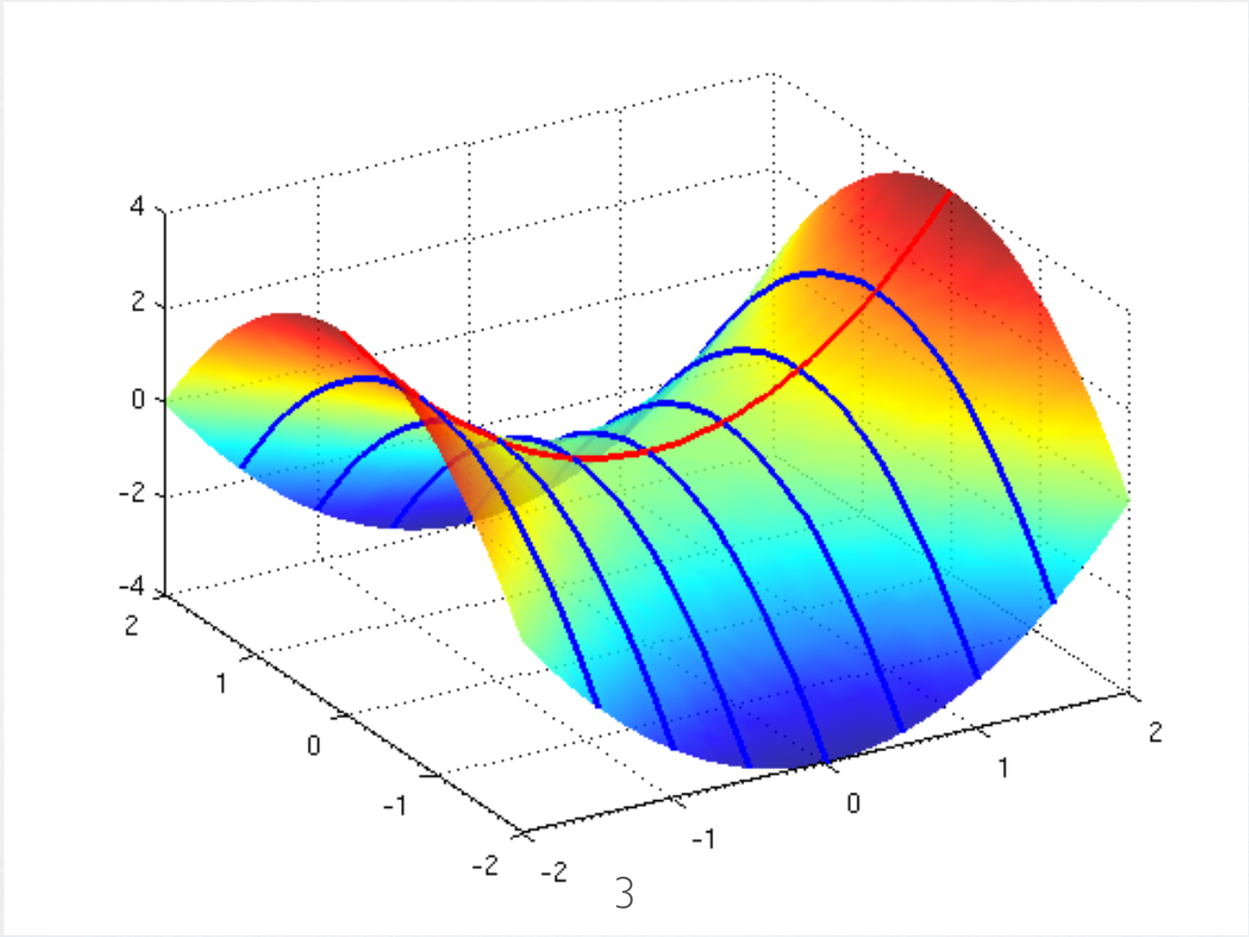
Simple case

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax + b = 0 \end{aligned}$$

"Saddle-point" form

$$\min_x \max_\lambda \underline{f(x) + \langle \lambda, Ax + b \rangle}$$

Lagrangian



LAGRANGIAN APPROACH

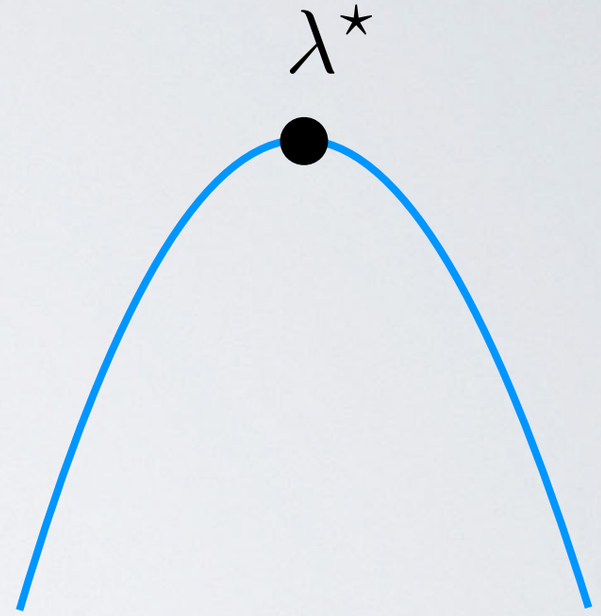
$$\text{minimize} \quad f(x) \quad \text{subject to} \quad Ax + b = 0$$

Dual approach: maximize dual function

$$d(\lambda) = \min_x f(x) + \langle \lambda, Ax + b \rangle$$

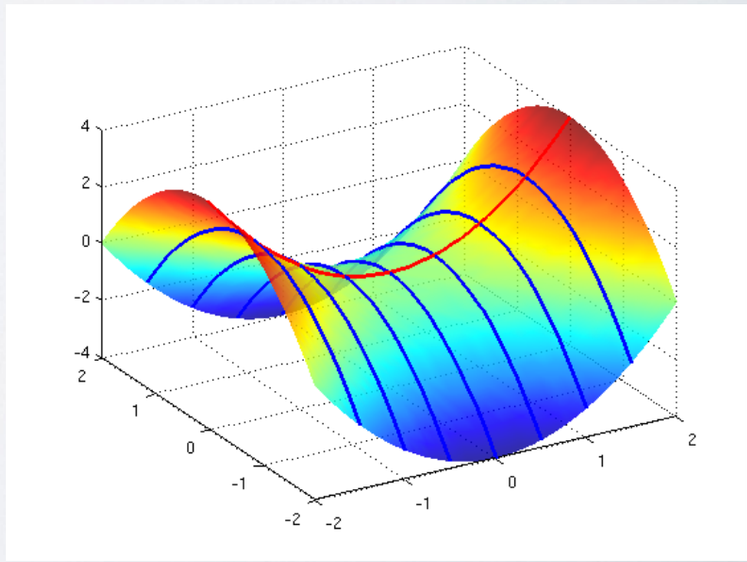
using conjugate

$$d(\lambda) = \langle b, \lambda \rangle - f^*(-A^T \lambda)$$



Lagrangian approach: directly find saddle point of Lagrangian

$$\max_\lambda \min_x f(x) + \langle \lambda, Ax + b \rangle$$



UZAWA'S METHOD

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad Ax + b = 0$$

$$\max_{\lambda} \min_{x} f(x) + \langle \lambda, Ax + b \rangle$$

Uzawa's method

minimize $x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax + b \rangle$

gradient ascent $\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + b)$

UZAWA'S METHOD

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad Ax + b = 0$$

Uzawa's method

minimize $\quad x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax + b \rangle$

gradient ascent $\quad \lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + b)$

dual

$$d(\lambda) = \langle b, \lambda \rangle - f^*(-A^T\lambda)$$

optimality

$$0 \in \partial f(x^{k+1}) + A^T\lambda^k$$

$$-A^T\lambda^k \in \partial f(x^{k+1})$$

$$x^{k+1} \in \partial f^*(-A^T\lambda^k)$$

$$Ax^{k+1} + b \in A\partial f^*(-A^T\lambda^k) + b$$

dual gradient

GRADIENT ASCENT

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad Ax + b = 0$$

Uzawa's method

minimize $x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax + b \rangle$

gradient ascent $\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + b)$

dual

$$d(\lambda) = \langle b, \lambda \rangle - f^*(-A^T\lambda)$$

$$Ax^{k+1} + b \in \underbrace{A\partial f^*(-A^T\lambda^k) + b}_{\text{dual gradient}}$$

$$\lambda^{k+1} = \lambda^k + \tau \partial d(\lambda^k)$$

stepsize restriction?

CONVERGENCE

$$Ax^{k+1} + b \in \underbrace{A\partial f^*(-A^T\lambda^k)}_{\text{dual gradient}} + b$$

Nifty Theorem: Strong Convexity = Smooth Dual

If f is strongly convex with constant m,
then f^* has Lipschitz continuous gradient
with constant $L = 1/m$.

gradient ascent $\lambda^{k+1} = \lambda^k + \tau(Ax + b)$

$$L_{dual} = \|A^T A\| L_{f^*} = \frac{\|A^T A\|_{op}}{m}$$

$$\tau \leq 2/L_{dual} = \frac{2m}{\|A^T A\|_{op}}$$

Problem: requires **strong** convexity!

AUGMENTED LAGRANGIAN

idea: add curvature to the primal problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax + b = 0 \end{aligned}$$

Lagrangian

$$L(x, \lambda) = f(x) + \langle \lambda, Ax + b \rangle$$

adds curvature

Augmented Lagrangian

$$L_\tau(x, \lambda) = f(x) + \langle \lambda, Ax + b \rangle + \frac{\tau}{2} \|Ax + b\|^2$$

- optimality for λ: $Ax + b = 0$
- reduced energy: $f(x)$
- saddle point = solution to constrained problem

METHOD OF MULTIPLIERS

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax + b = 0 \end{aligned}$$

Augmented Lagrangian

$$\min_x \max_\lambda f(x) + \langle \lambda, Ax + b \rangle + \frac{\tau}{2} \|Ax + b\|^2$$

Method of Multipliers

minimize $x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax + b \rangle + \frac{\tau}{2} \|Ax + b\|^2$

gradient step $\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + b)$

WHY IS THIS BETTER?

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax + b = 0 \end{aligned}$$

Method of Multipliers

minimize $$x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax + b \rangle + \tfrac{\tau}{2} \|Ax + b\|^2$$

gradient step $$\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + b)$$

dual

$$d(\lambda) = \langle \lambda, b \rangle - f^*(-A\lambda)$$

$$0 \in \partial f(x^{k+1}) + A^T \lambda^k + \tau A^T (Ax^{k+1} + b)$$

$$0 \in \partial f(x^{k+1}) + A^T (\lambda^k + \tau(Ax^{k+1} + b))$$

$$0 \in \partial f(x^{k+1}) + A^T \lambda^{k+1}$$

WHY IS THIS BETTER?

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax + b = 0 \end{aligned}$$

dual

$$d(\lambda) = \langle \lambda, b \rangle - f^*(-A^T\lambda)$$

iterates satisfy…

$$0 \in \partial f(x^{k+1}) + A^T\lambda^{k+1}$$

$$-A^T\lambda^{k+1} \in \partial f(x^{k+1})$$

$$x^{k+1} \in \partial f^*(-A^T\lambda^{k+1})$$

$$Ax^{k+1} + b \in \underline{A\partial f^*(-A^T\lambda^{k+1}) + b}$$

dual gradient

MM=BACKWARD GRADIENT

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax + b = 0 \end{aligned}$$

dual

$$d(\lambda) = \langle \lambda, b \rangle - f^*(-A^T \lambda)$$

method of multipliers

$$\text{minimize} \quad x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax + b \rangle + \tfrac{\tau}{2} \|Ax + b\|^2$$

$$\text{gradient step} \quad \lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + b)$$

$$Ax^{k+1} + b \in \underbrace{A\partial f^*(-A^T \lambda^{k+1}) + b}_{\text{dual gradient}}$$

backward gradient ascent

$$\lambda^{k+1} = \lambda^k + \tau \partial d(\lambda^{k+1})$$

CONVERGENCE

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax + b = 0 \end{aligned}$$

backward gradient ascent

$$\lambda^{k+1} = \lambda^k + \tau \partial d(\lambda^{k+1})$$

Works for any stepsize!

problem: requires solution of problem on every iteration

$$x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax + b \rangle + \frac{\tau}{2} \|Ax + b\|^2$$

Can we solve the whole problem in one shot?

SPLIT OBJECTIVE

just like we did for FBS…

$$\begin{aligned} \text{minimize} \quad & f(x) + g(y) \\ \text{subject to} \quad & Ax + By + c = 0 \end{aligned}$$

$$L_\tau(x, y, \lambda) = f(x) + g(y) + \langle \lambda, Ax + By + c \rangle + \frac{\tau}{2} \|Ax + By + c\|^2$$

method of multipliers

$$x^{k+1}, y^{k+1} = \operatorname*{arg\,min}_{x,y} f(x) + g(y) + \langle \lambda^k, Ax + By + c \rangle + \frac{\tau}{2} \|Ax + By + c\|^2$$

$$\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + By^{k+1} + c)$$

ADMM

$$\begin{aligned} \text{minimize} \quad & f(x) + g(y) \\ \text{subject to} \quad & Ax + By + c = 0 \end{aligned}$$

method of multipliers

$$x^{k+1}, y^{k+1} = \arg\min_{x,y} f(x) + g(y) + \langle \lambda^k, Ax + By + c \rangle + \frac{\tau}{2} \|Ax + By + c\|^2$$

$$\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + By^{k+1} + c)$$

alternating direction method of multipliers

$$x^{k+1} = \arg\min_{x} f(x) + \langle \lambda^k, Ax \rangle + \frac{\tau}{2} \|Ax + By^k + c\|^2$$

$$y^{k+1} = \arg\min_{y} g(y) + \langle \lambda^k, By \rangle + \frac{\tau}{2} \|Ax^{k+1} + By + c\|^2$$

$$\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + By^{k+1} + c)$$

RESIDUALS

Lagrangian

$$L(x, y, \lambda) = f(x) + g(y) + \langle \lambda, Ax + By + c \rangle$$

How do we measure closeness to a saddle point?

optimality conditions

x-residual: $0 \in \partial f(x) + A^T \lambda$

y-residual: $0 \in \partial g(x) + B^T \lambda$

λ-residual: $0 = Ax + By + c$

RESIDUALS

Lagrangian

$$L(x, y, \lambda) = f(x) + g(y) + \langle \lambda, Ax + By + c \rangle$$

alternating direction method of multipliers

$$x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax \rangle + \frac{\tau}{2} \|Ax + By^k + c\|^2$$

$$y^{k+1} = \arg\min_y g(y) + \langle \lambda^k, By \rangle + \frac{\tau}{2} \|Ax^{k+1} + By + c\|^2$$

$$\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + By^{k+1} + c)$$

$$0 \in \partial g(y^{k+1}) + B^T \lambda^k + \tau B^T (Ax^{k+1} + By^{k+1} + c)$$

$$0 \in \partial g(y^{k+1}) + B^T (\lambda^k + \tau(Ax^{k+1} + By^{k+1} + c))$$

y-residual: $0 \in \partial g(y^{k+1}) + B^T \lambda^{k+1}$

RESIDUALS

Lagrangian

$$L(x, y, \lambda) = f(x) + g(y) + \langle \lambda, Ax + By + c \rangle$$

alternating direction method of multipliers

$$x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax \rangle + \frac{\tau}{2} \|Ax + By^k + c\|^2$$

$$y^{k+1} = \arg\min_y g(y) + \langle \lambda^k, By \rangle + \frac{\tau}{2} \|Ax^{k+1} + By + c\|^2$$

$$\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + By^{k+1} + c)$$

$$0 \in \partial f(x^{k+1}) + A^T(\lambda^k + \tau(Ax^{k+1} + By^k + c))$$

$$0 \in \partial f(x^{k+1}) + A^T(\lambda^k + \tau(Ax^{k+1} + By^{k+1} + c)) + \tau A^T B y^k - \tau A^T B y^{k+1}$$

$$0 \in \partial f(x^{k+1}) + A^T \lambda^{k+1} + \tau A^T B y^k - \tau A^T B y^{k+1}$$

x-residual: $\tau A^T B y^{k+1} - \tau A^T B y^k \in \partial f(x^{k+1}) + A^T \lambda^{k+1}$

RESIDUALS

Lagrangian

$$L(x, y, \lambda) = f(x) + g(y) + \langle \lambda, Ax + By + c \rangle$$

alternating direction method of multipliers

$$x^{k+1} = \arg\min_x f(x) + \langle \lambda^k, Ax \rangle + \frac{\tau}{2} \|Ax + By^k + c\|^2$$

$$y^{k+1} = \arg\min_y g(y) + \langle \lambda^k, By \rangle + \frac{\tau}{2} \|Ax^{k+1} + By + c\|^2$$

$$\lambda^{k+1} = \lambda^k + \tau(Ax^{k+1} + By^{k+1} + c)$$

lambda-residual: $Ax^{k+1} + By^{k+1} + c$

CONVERGENCE

Lagrangian

$$L(x, y, \lambda) = f(x) + g(y) + \langle \lambda, Ax + By + c \rangle$$

x (primal) residual: $\tau A^T B y^{k+1} - \tau A^T B y^k$

lambda (dual) residual: $Ax^{k+1} + By^{k+1} + c$

Theorem (He and Yuan '12)

For any fixed stepsize τ, ADMM converges in the residuals with rate

$$\|\tau A^T B y^{k+1} - \tau A^T B y^k\|^2 + \|Ax^{k+1} + By^{k+1} + c\|^2 < O\left(\frac{1}{k}\right)$$

LASSO

$$\text{minimize} \quad \mu|\underbrace{x}_{y}| + \frac{1}{2}\|Ax - b\|^2$$

"split Bregman" form

$$\text{minimize} \quad \mu|y| + \frac{1}{2}\|Ax - b\|^2$$

$$\text{subject to} \quad x - y = 0$$

augmented Lagrangian

$$\mu|y| + \frac{1}{2}\|Ax - b\|^2 + \langle\lambda, x - y\rangle + \frac{\tau}{2}\|x - y\|^2$$

LASSO

$$\text{minimize} \quad \mu|x| + \frac{1}{2}\|Ax - b\|^2$$

augmented Lagrangian

$$\mu|y| + \frac{1}{2}\|Ax - b\|^2 + \langle \lambda, x - y \rangle + \frac{\tau}{2}\|x - y\|^2$$

ADMM lasso

$$x^{k+1} = \arg\min_x \frac{1}{2}\|Ax - b\|^2 + \langle \lambda^k, x \rangle + \frac{\tau}{2}\|x - y^k\|^2$$

$$y^{k+1} = \arg\min_y \mu|y| - \langle \lambda^k, y \rangle + \frac{\tau}{2}\|x^{k+1} - y\|^2$$

$$\lambda^{k+1} = \lambda^k + \tau(x^{k+1} - y^{k+1})$$

how do you solve these sub-problems?

EXAMPLE: SPLIT BREGMAN

$$\text{minimize} \quad \mu|\nabla x| + \frac{1}{2}\|Ax - f\|^2$$

"split Bregman" form

$$\text{minimize} \quad \mu|y| + \frac{1}{2}\|Ax - f\|^2$$

$$\text{subject to} \quad \nabla x - y = 0$$

augmented Lagrangian

$$\mu|y| + \frac{1}{2}\|Ax - f\|^2 + \langle \lambda, \nabla x - y \rangle + \frac{\tau}{2}\|\nabla x - y\|^2$$

EXAMPLE: SPLIT BREGMAN

$$\text{minimize} \quad \mu|\nabla x| + \frac{1}{2}\|Ax - f\|^2$$

$$\mu|y| + \frac{1}{2}\|Ax - f\|^2 + \langle \lambda, \nabla x - y \rangle + \frac{\tau}{2}\|\nabla x - y\|^2$$

Split Bregman TV

$$x^{k+1} = \arg\min_x \frac{1}{2}\|Ax - f\|^2 + \langle \lambda^k, \nabla x \rangle + \frac{\tau}{2}\|\nabla x - y^k\|^2$$

$$y^{k+1} = \arg\min_y \mu|y| - \langle \lambda^k, y \rangle + \frac{\tau}{2}\|\nabla x^{k+1} - y\|^2$$

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

X-UPDATE

$$\text{minimize} \quad \mu|\nabla x| + \frac{1}{2}\|Ax - f\|^2$$

$$\mu|y| + \frac{1}{2}\|Ax - f\|^2 + \langle \lambda, \nabla x - y \rangle + \frac{\tau}{2}\|\nabla x - y\|^2$$

x-update

$$x^{k+1} = \arg\min_x \frac{1}{2}\|Ax - f\|^2 + \langle \lambda^k, \nabla x \rangle + \frac{\tau}{2}\|\nabla x - y^k\|^2$$

optimality condition

$$A^T(Ax - f) + \nabla^T \lambda^k + \tau \nabla^T (\nabla x - y^k) = 0$$

linear system

$$(A^T A + \tau \nabla^T \nabla)x = A^T f - \nabla^T \lambda^k + \tau \nabla^T y^k$$

X-UPDATE

$$\text{minimize} \quad \mu|\nabla x| + \frac{1}{2}\|Ax - f\|^2$$

linear system

$$(A^T A + \tau \nabla^T \nabla)x = A^T f - \nabla^T \lambda^k + \tau \nabla^T y^k$$

deblurring: $A = F^H D F$

$$(F^T D^H D F + \tau \nabla^T \nabla)x = rhs$$

$$(F^T |D|^2 F + \tau \nabla^T \nabla)x = rhs$$

$\nabla = F^H K F$

$$(F^T |D|^2 F + \tau F^T |K|^2 F)x = rhs$$

$$F^T (|D|^2 + \tau |K|^2) F x = rhs$$

$$x^{k+1} = F^T (|D|^2 + \tau |K|^2)^{-1} F(rhs)$$

Y-UPDATE

$$\text{minimize} \quad \mu|\nabla x| + \frac{1}{2}\|Ax - f\|^2$$

$$\mu|y| + \frac{1}{2}\|Ax - f\|^2 + \langle \lambda, \nabla x - y \rangle + \frac{\tau}{2}\|\nabla x - y\|^2$$

y-update

$$y^{k+1} = \arg\min_y \mu|y| - \langle \lambda^k, y \rangle + \frac{\tau}{2}\|\nabla x^{k+1} - y\|^2$$

complete square

$$y^{k+1} = \arg\min_y \mu|y| + \frac{\tau}{2}\|y - \nabla x^{k+1} - \frac{1}{\tau}\lambda^k\|^2$$

proximal operator

$$y^{k+1} = \text{shrink}(\nabla x^{k+1} + \frac{1}{\tau}\lambda^k, \mu/\tau)$$

DEBLURRING ALGORITHM

$$\text{minimize} \quad \mu|\nabla x| + \frac{1}{2}\|Ax - f\|^2$$

$$\mu|y| + \frac{1}{2}\|Ax - f\|^2 + \langle \lambda, \nabla x - y \rangle + \frac{\tau}{2}\|\nabla x - y\|^2$$

Split Bregman Deblurring

$$x^{k+1} = F^H(|D|^2 + \tau|K|^2)^{-1}F(F^H D^H F f - \nabla^T \lambda^k + \tau \nabla^T y^k)$$

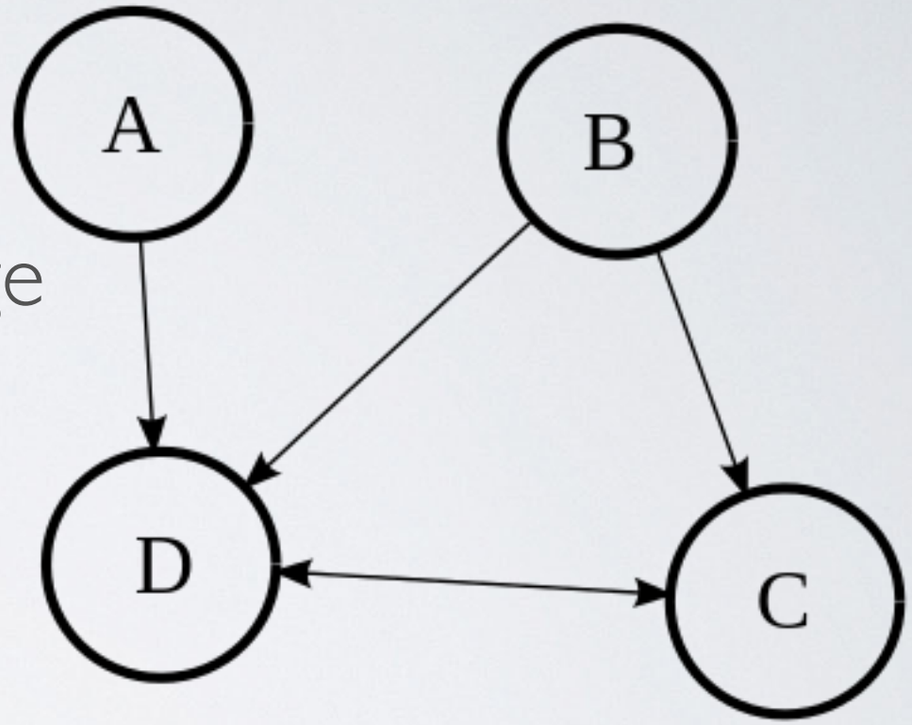
$$y^{k+1} = \text{shrink}(\nabla x^{k+1} + \frac{1}{\tau}\lambda^k, \mu/\tau)$$

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

INVERSE COVARIANCE SELECTION

Graphical model:

- nodes are random variables
- **dependent** variables connected by edge

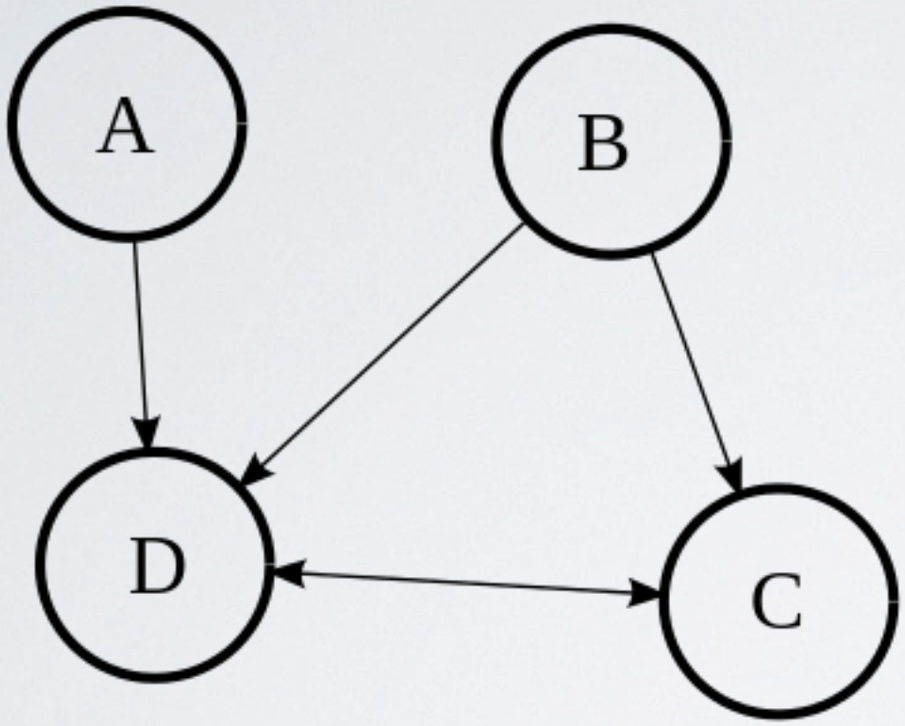


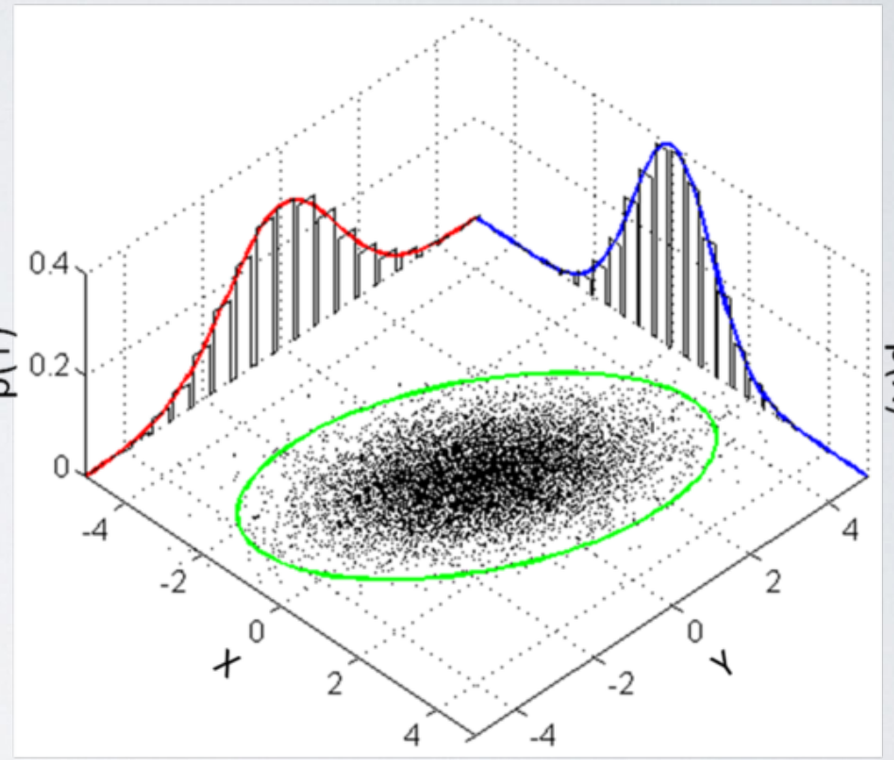
Example: "gene finding"

- measure mRNA expression levels
- try to find mRNA strands regulated by same genes

Variables are **conditionally independent** if there is no path from one to another

INVERSE COVARIANCE SELECTION





$$x_i = \begin{pmatrix} a_i \\ b_i \\ c_i \\ d_i \end{pmatrix}$$

inverse covariance matrix

$$p(x) = \frac{1}{(2\pi)^p |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2} x^T \Sigma^{-1} x)$$

LIKELIHOOD MODEL

$$p(x) = \frac{1}{(2\pi)^p |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2} x^T \Sigma^{-1} x)$$

$$l(\Sigma) = \prod_i \frac{1}{(2\pi)^p |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2} x_i^T \Sigma^{-1} x_i)$$

$$ll(\Sigma) = -np \log(2\pi) - \frac{n}{2} \log\det\Sigma - \frac{1}{2} \sum_i x_i^T \Sigma^{-1} x_i$$

multiply by -1 to make it a minimization problem

$$\log\det\Sigma + \frac{1}{n} \sum_i x_i^T \Sigma^{-1} x_i$$

LIKELIHOOD MODEL

$$\log\det\Sigma + \frac{1}{n}\sum_i x_i^T\Sigma^{-1}x_i$$

$$\frac{1}{n}\sum_i x_i^T\Sigma^{-1}x_i = \frac{1}{n}\sum_i \langle x_i x_i^T, \Sigma^{-1}\rangle = \langle \frac{1}{n}\sum_i x_i x_i^T, \Sigma^{-1}\rangle = \langle S, \Sigma^{-1}\rangle$$

negative log likelihood

$$\log\det\Sigma + \langle S, \Sigma^{-1}\rangle$$

empirical covariance

sparse model

$$\underset{\Sigma^{-1}}{\text{minimize}} \quad \log\det\Sigma + \langle S, \Sigma^{-1}\rangle + |\Sigma^{-1}|$$

change variables

$$\underset{X}{\text{minimize}} \quad -\log\det X + \langle S, X\rangle + |X|$$

ADMM FOR SICS

$$\underset{X}{\text{minimize}} \quad -\log\det X + \langle S, X\rangle + |X|$$

split Bregman form

$$\underset{X}{\text{minimize}} \quad -\log\det X + \langle S, X\rangle + |Y|$$
$$\text{subject to} \;\; X = Y$$

augmented Lagrangian

$$\max_{\lambda}\min_{X,Y} -\log\det X + \langle S, X\rangle + |Y| + \langle \lambda, X - Y\rangle + \frac{\tau}{2}\|X - Y\|^2$$

ADMM FOR SICS

augmented Lagrangian

$$\max_{\lambda} \min_{X,Y} -\log\det X + \langle S, X\rangle + |Y| + \langle \lambda, X - Y\rangle + \frac{\tau}{2}\|X - Y\|^2$$

step 1: minimize for Y

$$\underset{Y}{\text{minimize}} \quad |Y| + \langle \lambda, X - Y\rangle + \frac{\tau}{2}\|Y - X\|^2$$

$$\underset{Y}{\text{minimize}} \quad |Y| + \frac{\tau}{2}\|Y - X - \frac{1}{\tau}\lambda\|^2$$

solution

$$Y^{k+1} = \text{shrink}(X^k + \tau^{-1}\lambda^k)$$

ADMM FOR SICS

augmented Lagrangian

$$\max_{\lambda} \min_{X,Y} -\log\det X + \langle S, X\rangle + |Y| + \langle \lambda, X - Y\rangle + \frac{\tau}{2}\|X - Y\|^2$$

step 2: minimize for X

$$\underset{X}{\text{minimize}} \quad -\log\det X + \langle S, X\rangle + \langle \lambda, X\rangle + \frac{\tau}{2}\|X - Y\|^2$$

$$\underset{X}{\text{minimize}} \quad -\log\det X + \frac{\tau}{2}\|X - \underbrace{Y + \tau^{-1}(\lambda + S)}\|^2$$

$$X = U\Lambda_X U^T \qquad\qquad Z = U\Lambda_Z U^T$$

ADMM FOR SICS

step 2: minimize for X

$$\underset{X}{\text{minimize}} \quad -\log\det X + \langle S, X\rangle + \langle \lambda, X\rangle + \frac{\tau}{2}\|X - Y\|^2$$

$$\underset{X}{\text{minimize}} \quad -\log\det X + \frac{\tau}{2}\|X - Y + \tau^{-1}(\lambda + S)\|^2$$

$$X = U\Lambda_X U^T \qquad\qquad Z = U\Lambda_Z U^T$$

$$\sum_i -\log\sigma_X^i + \frac{\tau}{2}(\sigma_X^i - \sigma_Z^i)^2 \longrightarrow \frac{-1}{\sigma_X^i} + \tau(\sigma_X^i - \sigma_Z^i) = 0$$

$$\sigma_X^i = \frac{\sigma_Z^i + \sqrt{(\sigma_Z^i)^2 + 4/\tau}}{2}$$

final step: $\lambda^{k+1} = \lambda^k + \tau(X - Y)$

SCALED ADMM

$$\begin{aligned} \text{minimize} \quad & f(x) + g(y) \\ \text{subject to} \quad & Ax + By + c = 0 \end{aligned}$$

augmented Lagrangian

$$L_\tau(x, y, \lambda) = f(x) + g(y) + \langle \lambda, Ax + By + c \rangle + \frac{\tau}{2} \|Ax + By + c\|^2$$

scaled Lagrangian

$$L_\tau(x, y, \lambda) = f(x) + g(y) + \frac{\tau}{2} \|Ax + By + c + \frac{1}{\tau}\lambda\|^2$$

These differ by a constant. Why??

penalty function/spring interpretation

SCALED ADMM

$$\begin{aligned} \text{minimize} \quad & f(x) + g(y) \\ \text{subject to} \quad & Ax + By + c = 0 \end{aligned}$$

scaled Lagrangian

$$L_\tau(x, y, \lambda) = f(x) + g(y) + \frac{\tau}{2}\|Ax + By + c + \frac{1}{\tau}\lambda\|^2$$

$$\lambda \leftarrow \lambda$$

$$L_\tau(x, y, \hat{\lambda}) = f(x) + g(y) + \frac{\tau}{2}\|Ax + By + c + \hat{\lambda}\|^2$$

scaled ADMM

$$x^{k+1} = \arg\min_x f(x) + \frac{\tau}{2}\|Ax + By^k + c + \hat{\lambda}^k\|^2$$

$$y^{k+1} = \arg\min_y g(y) + \frac{\tau}{2}\|Ax^{k+1} + By + c + \hat{\lambda}^k\|^2$$

$$\hat{\lambda}^{k+1} = \hat{\lambda}^k + Ax^{k+1} + By^{k+1} + c$$

DISTRIBUTED PROBLEMS

$$\text{minimize} \quad g(x) + \sum_i f_i(x)$$

example: sparse least squares

$$\text{minimize} \quad \mu|x| + \frac{1}{2}\|Ax - b\|^2$$

$$\text{minimize} \quad \mu|x| + \sum_i \frac{1}{2}\|A_i x - b_i\|^2$$

$$A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_N \end{pmatrix}$$

data stored on different servers

CONSENSUS ADMM

$$\text{minimize} \quad g(x) + \sum_i f_i(x)$$

Central server holds global variables: z

Every client gets local copy of unknowns: x_i

$$\text{minimize} \quad g(z) + \sum_i f_i(x_i)$$

$$\text{subject to} \quad x_i = z, \ \forall i$$

CONSENSUS ADMM

$$\text{minimize} \quad g(z) + \sum_i f_i(x_i)$$

$$\text{subject to} \quad x_i = z, \ \forall i$$

scaled augmented Lagrangian

$$L = g(z) + \sum_i f_i(x_i) + \sum_i \frac{\tau}{2} \|x_i - z + \lambda_i\|^2$$

consensus ADMM

$$\text{central server:} \quad z^{k+1} = \arg\min_z g(z) + \sum_i \frac{\tau}{2} \|x_i^k - z + \lambda_i^k\|^2$$

$$\text{remote client:} \quad x_i^{k+1} = \arg\min_{x_i} f_i(x_i) + \frac{\tau}{2} \|x_i - z^{k+1} + \lambda_i^k\|^2$$

$$\text{remote client:} \quad \lambda_i^{k+1} = \lambda_i^k + x_i^{k+1} - z^{k+1}$$

CENTRAL STEP

$$\text{central server: } z^{k+1} = \arg\min_z g(z) + \sum_i \frac{\tau}{2}\|x_i^k - z + \lambda_i^k\|^2$$

$$z^{k+1} = \arg\min_z g(z) + \sum_i \frac{\tau}{2}\|z - (x_i^k + \lambda_i^k)\|^2$$

$$z^{k+1} = \arg\min_z g(z) + \frac{N\tau}{2}\|z - \underline{\frac{1}{N}\sum_i (x_i^k + \lambda_i^k)}\|^2$$

average of remote values

η^k

$$z^{k+1} = \arg\min_z g(z) + \frac{N\tau}{2}\|z - \eta^k\|^2$$

EXAMPLE: LASSO

$$\text{minimize} \quad \mu|x| + \sum_i \frac{1}{2}\|A_i x - b_i\|^2$$

scaled augmented Lagrangian

$$L = \mu|z| + \sum_i \frac{1}{2}\|A_i x_i - b_i\|^2 + \sum_i \frac{\tau}{2}\|x_i - z + \lambda_i\|^2$$

consensus LASSO

$$\text{MPI reduce: } \eta^k = \frac{1}{N}\sum_i x_i^k + \lambda_i^k$$

$$\text{central server: } z^{k+1} = \arg\min_z \mu|z| + \frac{N\tau}{2}\|z - \eta\|^2$$

$$\text{remote client: } x_i^{k+1} = \arg\min_{x_i} \frac{1}{2}\|A_i x_i - b_i\|^2 + \frac{\tau}{2}\|x_i - z^{k+1} + \lambda_i^k\|^2$$

$$\text{remote client: } \lambda_i^{k+1} = \lambda_i^k + x_i^{k+1} - z^{k+1}$$

WHY PDHG?

$$\text{minimize} \quad f(Ax) + g(x)$$

$$\text{minimize} \quad f(y) + g(x)$$
$$\text{subject to} \;\; Ax - y = 0$$

standard augmented Lagrangian

$$L(x, y, \lambda) = f(y) + g(x) + \langle \lambda, Ax - y \rangle + \frac{\tau}{2} \|Ax - y\|^2$$

how to minimize for x?

EXAMPLE: SPLIT BREGMAN

$$\text{minimize} \quad \mu|\nabla x| + \frac{1}{2}\|Ax - f\|^2$$

$$\mu|y| + \frac{1}{2}\|Ax - f\|^2 + \langle \lambda, \nabla x - y \rangle + \frac{\tau}{2}\|\nabla x - y\|^2$$

Split Bregman TV

$$x^{k+1} = \arg\min_x \frac{1}{2}\|Ax - f\|^2 + \langle \lambda^k, \nabla x \rangle + \frac{\tau}{2}\|\nabla x - y^k\|^2$$

$$y^{k+1} = \arg\min_y \mu|y| - \langle \lambda^k, y \rangle + \frac{\tau}{2}\|\nabla x^{k+1} - y\|^2$$

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

x-update

$$(A^T A + \tau \nabla^T \nabla)x = A^T f - \nabla^T \lambda^k + \tau \nabla^T y^k$$

PDHG: SADDLE-POINT PROBLEMS

$$\min_{x \in X} \max_{y \in Y} f(x) + \langle Ax, y \rangle - g(y)$$

- Convex functions: f and g
- "Link" term: $\langle Ax, y \rangle$
- We can evaluate "proximal operators"

$$J_{\tau F}(\hat{x}) = \operatorname*{argmin}_{x \in X} f(x) + \frac{1}{2\tau} \|x - \hat{x}\|^2$$

$$J_{\sigma G}(\hat{y}) = \operatorname*{argmin}_{y \in Y} g(y) + \frac{1}{2\sigma} \|y - \hat{y}\|^2.$$

HOW PDHG WORKS

$$\min_{x \in X} \max_{y \in Y} f(x) + \langle Ax, y \rangle - g(y)$$

Gradient descent
$$\hat{x} = x_k - \tau A^T y$$

Primal Proximal
$$x_{k+1} = \operatorname{argmin} f(x) + \frac{1}{2\tau} \|x - \hat{x}\|^2$$

Predict
$$\bar{x} = x_{k+1} + (x_{k+1} - x_k)$$

Gradient Ascent
$$\hat{y} = y_k + \sigma A \bar{x}$$

Dual Proximal
$$y_{k+1} = \operatorname{argmin} g(y) + \frac{1}{2\sigma} \|y - \hat{y}\|^2$$

FORMING THE SADDLE-POINT PROBLEM

$$\begin{aligned} \text{minimize} \quad & f(x) + h(Ax) \\ \downarrow & \\ \text{minimize} \quad & f(x) + h(y) \\ \text{subject to} \quad & Ax - y = 0 \\ \downarrow & \\ \text{minimize} \quad & f(x) + \langle \lambda, Ax \rangle - \langle \lambda, y \rangle + h(y) \end{aligned}$$

reminder

$$h^*(\lambda) = \max_y \langle \lambda, y \rangle - h(y) \quad \rightarrow \quad -h^*(\lambda) = \min_y -\langle \lambda, y \rangle + h(y)$$

$$\text{minimize} \quad f(x) + \langle \lambda, Ax \rangle - h^*(\lambda)$$

EXAMPLE: TV

$$\min_{x \in X} \max_{y \in Y} f(x) + \langle Ax, y \rangle - g(y)$$

- The problem: $|\nabla x| + \frac{\mu}{2} \|Ax - f\|^2$

Note: $|x| = \max_{|y| \leq 1} yx$

- Re-write TV: $|\nabla x| = \max_{\|y\|_\infty \leq 1} \langle y, \nabla x \rangle$

$$= \max_y \langle y, \nabla x \rangle - \chi_\infty(y)$$

- Saddle-Point form:

$$\max_y \min_x \ \frac{\mu}{2} \|Ax - f\|^2 + \langle y, \nabla x \rangle - \chi_\infty(y)$$

RESIDUALS

$$\min_{x \in X} \max_{y \in Y} f(x) + \langle Ax, y \rangle - g(y)$$

- Primal residual $p(x, y) = \partial f(x) + A^T y$
- Dual residual $d(x, y) = \partial g(y) - Ax$

$$x^{k+1} = \arg\min f(x) + \frac{1}{2\tau} \|x - \hat{x}\|^2$$

27

RESIDUALS

$$\min_{x \in X} \max_{y \in Y} f(x) + \langle Ax, y \rangle - g(y)$$

- Primal residual $p(x, y) = \partial f(x) + A^T y$
- Dual residual $d(x, y) = \partial g(y) - Ax$

$$x^{k+1} = \arg\min f(x) + \frac{1}{2\tau} \|x - \hat{x}\|^2$$

$$0 \in \partial f(x^{k+1}) + \frac{1}{\tau}(x^{k+1} - \hat{x})$$

$$= \partial f(x^{k+1}) + \frac{1}{\tau}(x^{k+1} - x^k) + A^T y^k$$

$$p(x, y) = \frac{1}{\tau}(x^k - x^{k+1}) - A^T(y^k - y^{k+1})$$

RESIDUALS

$$\min_{x \in X} \max_{y \in Y} f(x) + \langle Ax, y \rangle - g(y)$$

- Primal residual $p(x, y) = \partial f(x) + A^T y$
- Dual residual $d(x, y) = \partial g(y) - Ax$

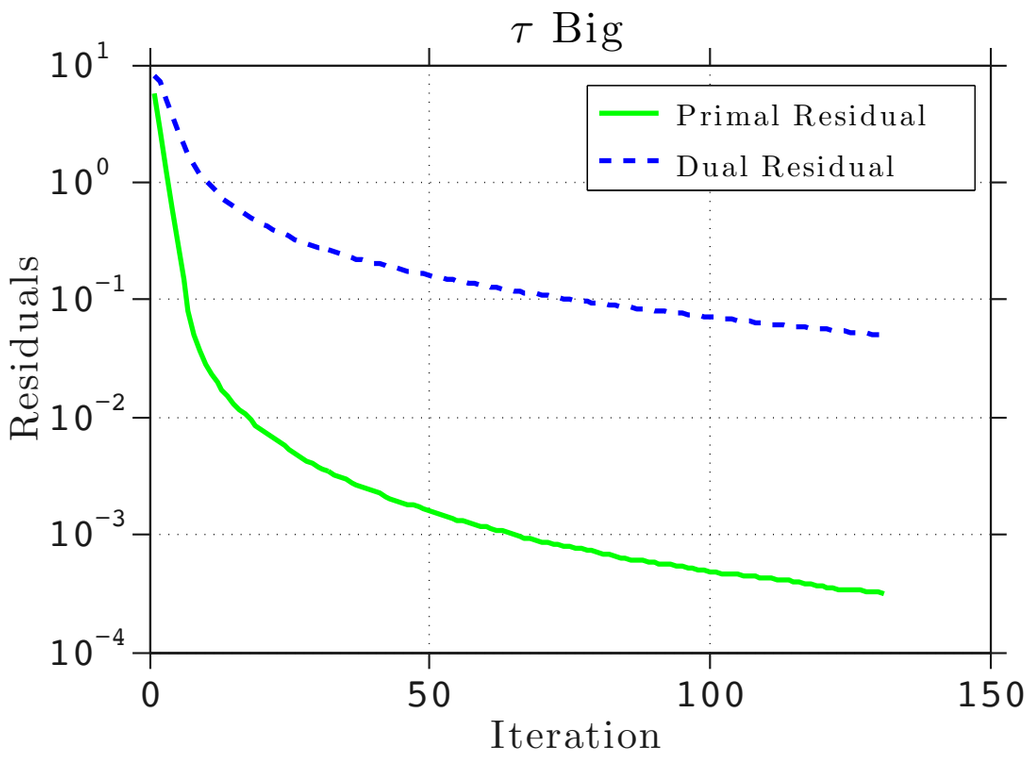
explicit formula

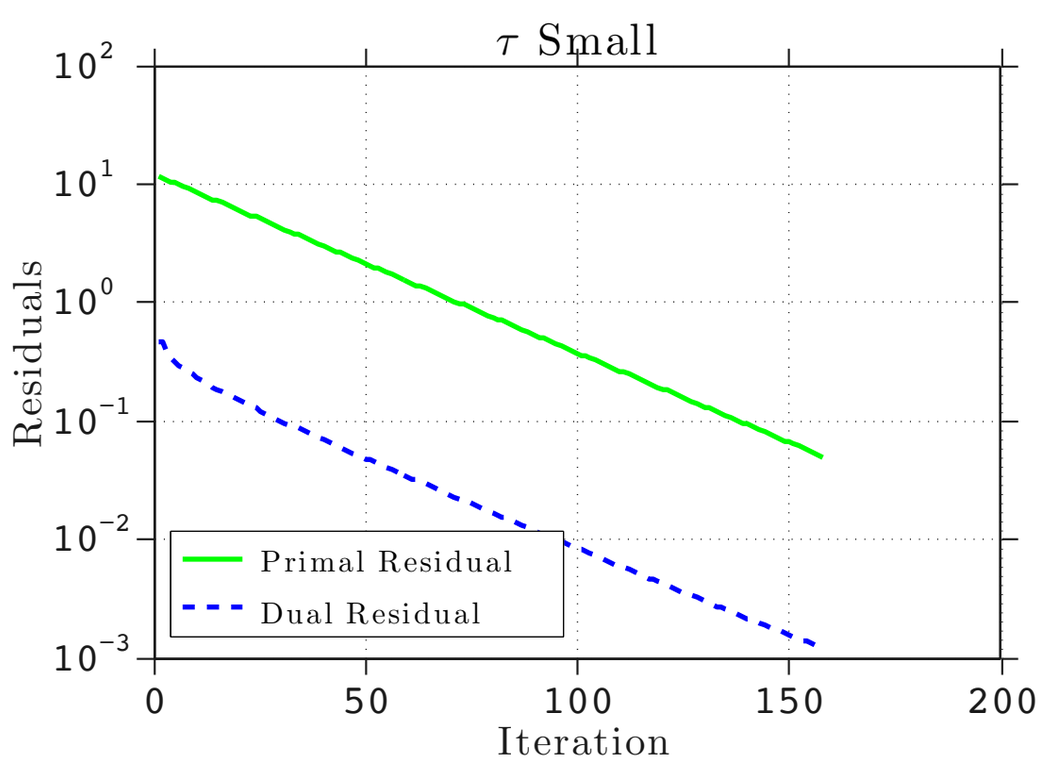
$$p(x, y) = \frac{1}{\tau}(x^k - x^{k+1}) - A^T(y^k - y^{k+1})$$

$$d(x, y) = \frac{1}{\sigma}(y^k - y^{k+1}) - A^T(x^k - x^{k+1})$$

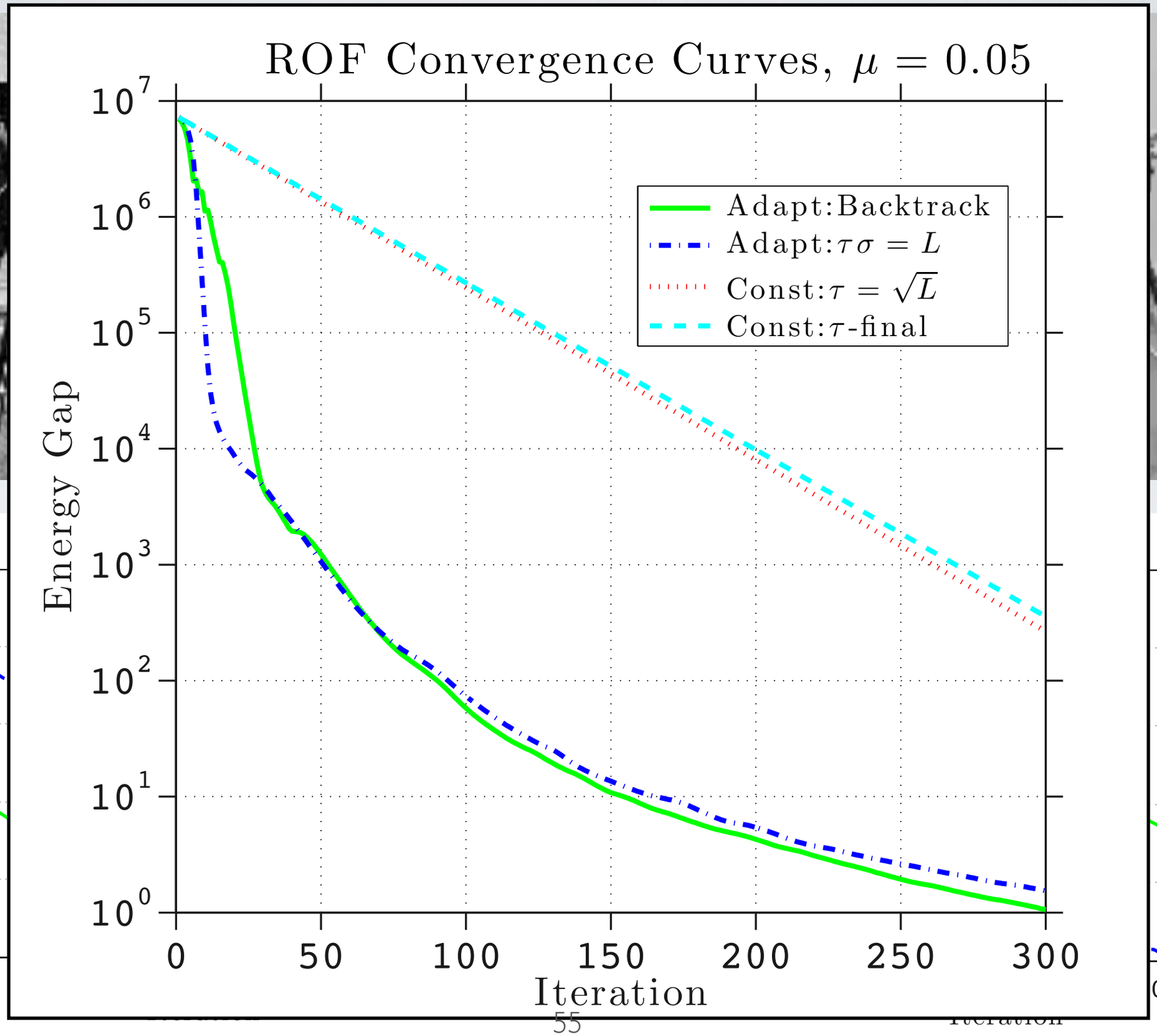
EXAMPLE: DENOISING

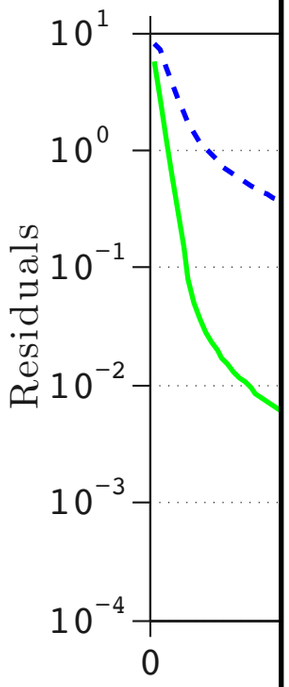
$$\min_x \frac{\mu}{2}\|x - f\|^2 + |\nabla x|$$

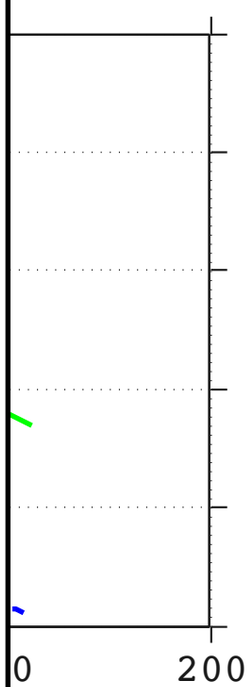




EXAMPLE: DENOISING







THEOREM

$$\phi_k = \max\left\{\frac{\tau_k - \tau_{k+1}}{\tau_k}, \frac{\sigma_k - \sigma_{k+1}}{\sigma_k}, 0\right\}$$

Theorem

PDHG converges if the following conditions hold:

A The stepsizes $\{\tau_k\}$ and $\{\sigma_k\}$ are bounded.

B The sequence $\{\phi_k\}$ is summable, i.e., $\sum_{k\geq 0} \phi_k < \infty$.

C One of the following holds:

C1 There is a constant L with

$$\tau_k \sigma_k < L < \rho(A^T A)^{-1}.$$

C2 Either X or Y is bounded, and

$$\frac{\gamma}{\tau_k}\|x_{k+1} - x_k\|^2 + \frac{\gamma}{\sigma_k}\|y_{k+1} - y_k\|^2 > 2\langle A(x_{k+1} - x_k), y_{k+1} - y_k\rangle.$$

BACKTRACKING

PDHG

Pick τ, σ with $\tau\sigma > \|A^T A\|^{-1}$ $\qquad \gamma = 0.75$

While p^k and d^k are "big:"

Do PDHG, get x^{k+1} and y^{k+1}

If $\frac{\gamma}{\tau}\|x^{k+1} - x_k\|^2 + \frac{\gamma}{\sigma}\|y^{k+1} - y_k\|^2 < 2\langle A(x^{k+1} - x_k), y^{k+1} - y_k\rangle$

$\tau \leftarrow \tau/2, \;\; \sigma \leftarrow \sigma/2,$

Redo PDHG, get x^{k+1} and y^{k+1}

BACKTRACKING

Pick τ, σ with $\tau\sigma > \|A^T A\|^{-1}$ $\qquad$ $\gamma = 0.75$

While p^k and d^k are "big:"

Do PDHG, get x^{k+1} and y^{k+1}

backtracking

If $\frac{\gamma}{\tau}\|x^{k+1} - x_k\|^2 + \frac{\gamma}{\sigma}\|y^{k+1} - y_k\|^2 < 2\langle A(x^{k+1} - x_k), y^{k+1} - y_k\rangle$

$\tau \leftarrow \tau/2, \;\; \sigma \leftarrow \sigma/2,$

Redo PDHG, get x^{k+1} and y^{k+1}

WHAT'S A GRAPH CUT

cut function

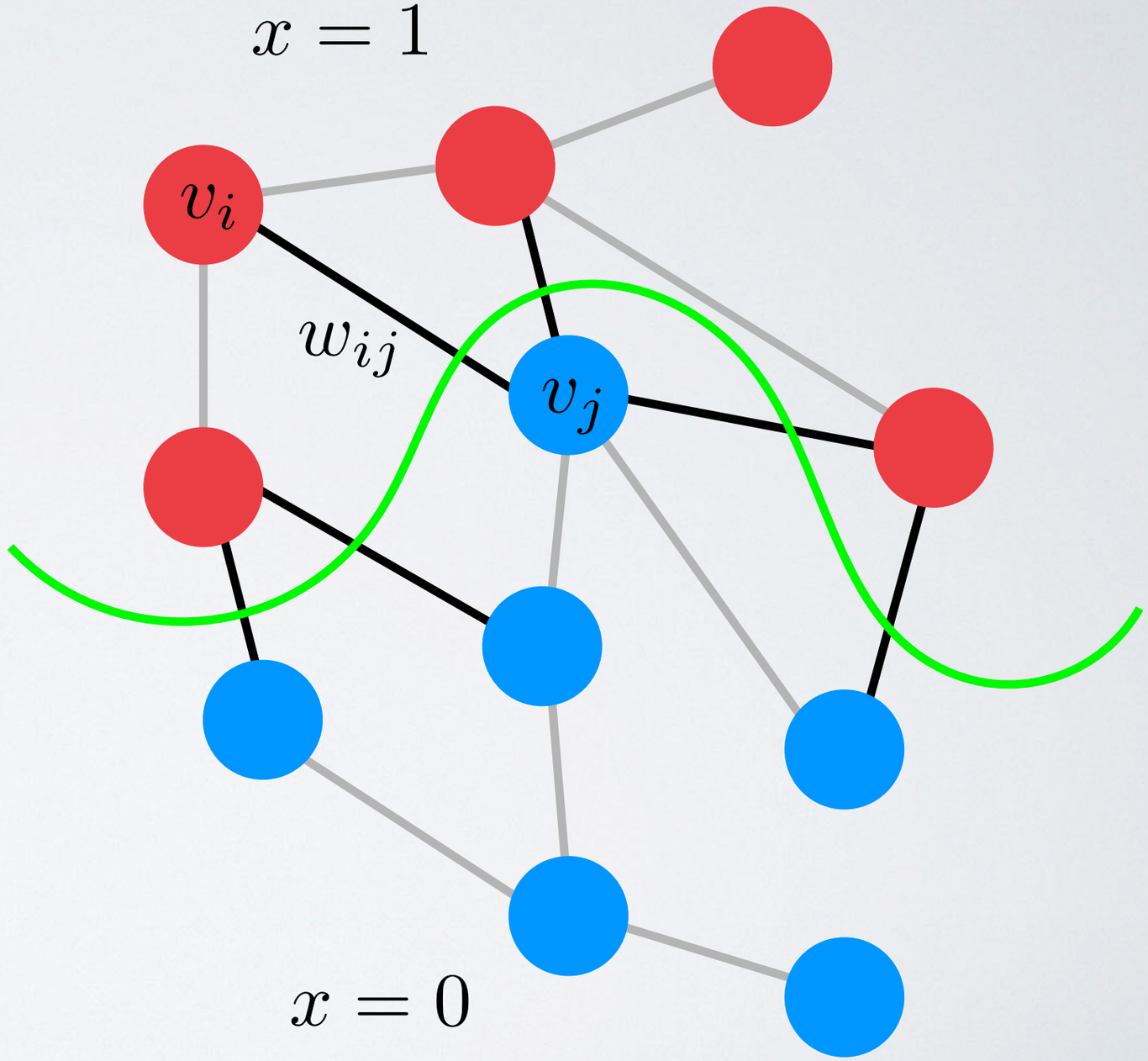
$$x : V \rightarrow \{0, 1\}$$

cut energy

$$\sum_{i,j} |x_i - x_j| w_{i,j} + \sum_i x_i f(i)$$

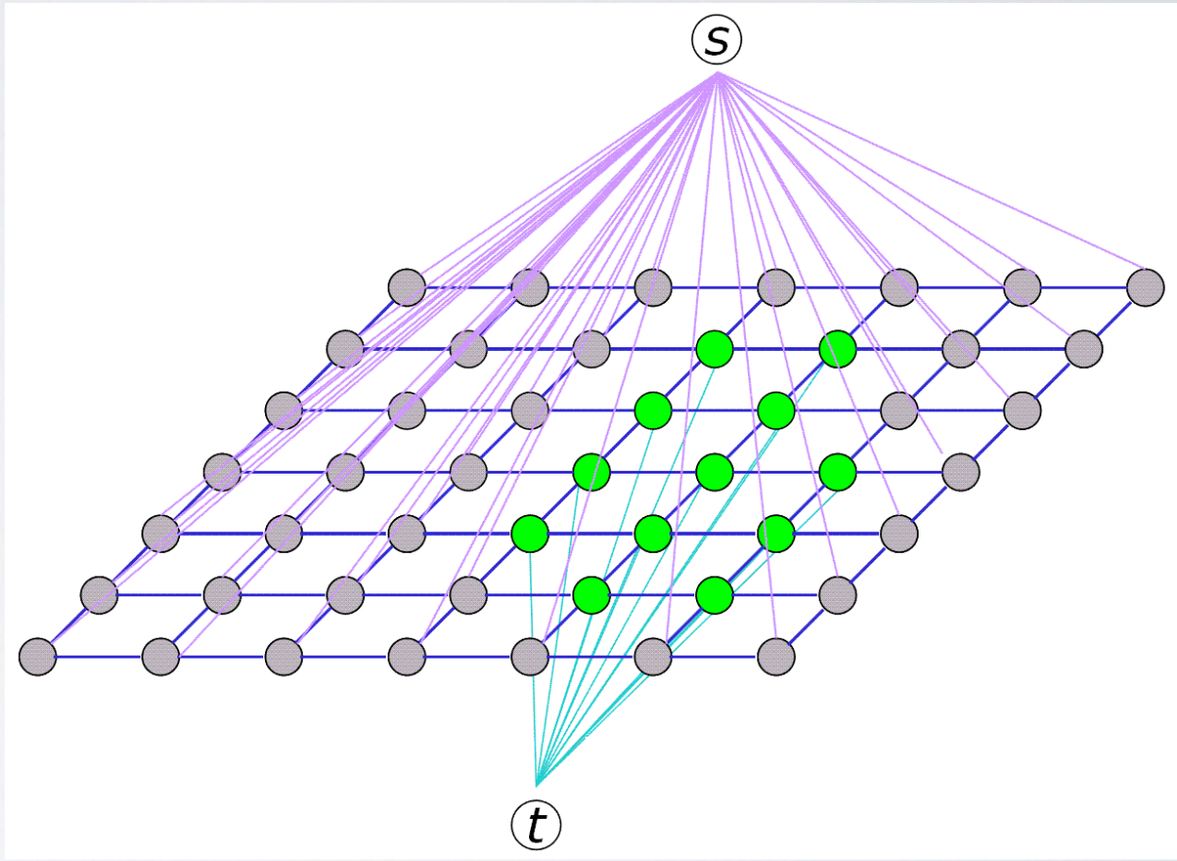
must be positive (why?)

can be anything



MIN CUT PROBLEM

Cut graph into parts such that s and t lie in different chunks



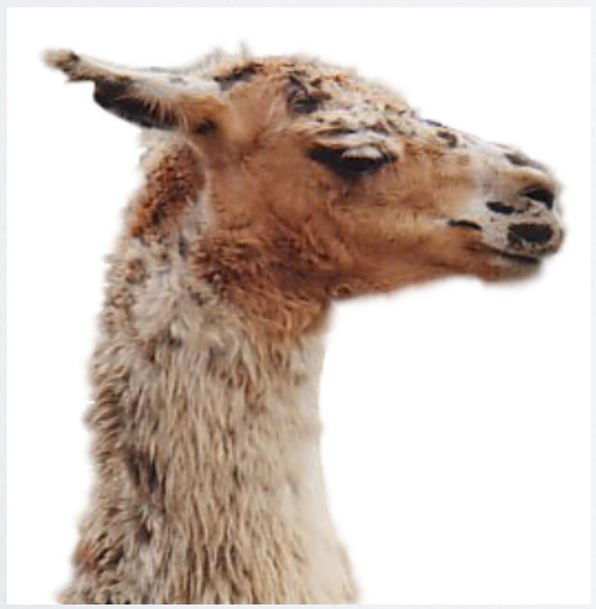
graph cuts solves this

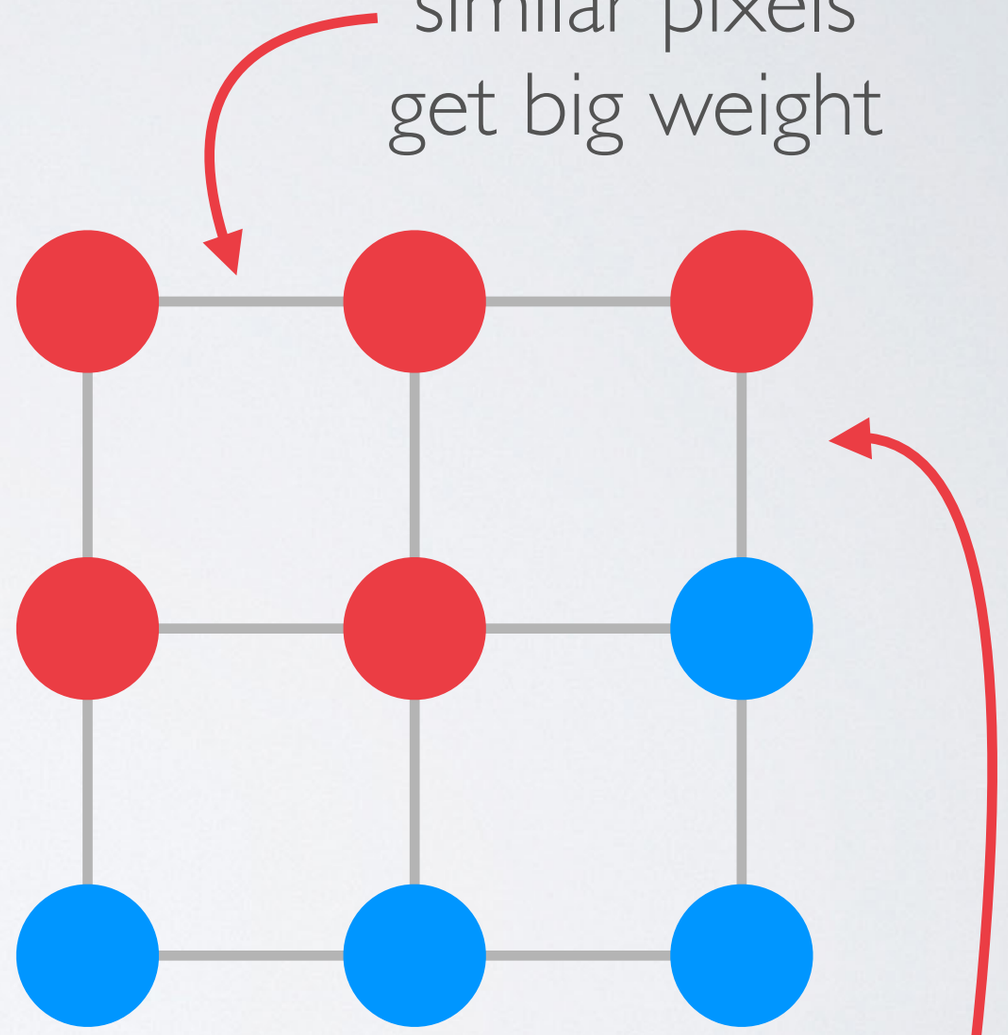
$$\underset{x}{\text{minimize}} \sum_{i,j} |x_i - x_j| w_{i,j} + \sum_i x_i f(i)$$

$$\text{subject to} \quad x_s = 1, x_t = 0$$

EXAMPLE: SEGMENTATION

$$w_{ij} = \frac{e^{-\|v_i - v_j\|^2/\sigma^2}}{d(i,j)}$$

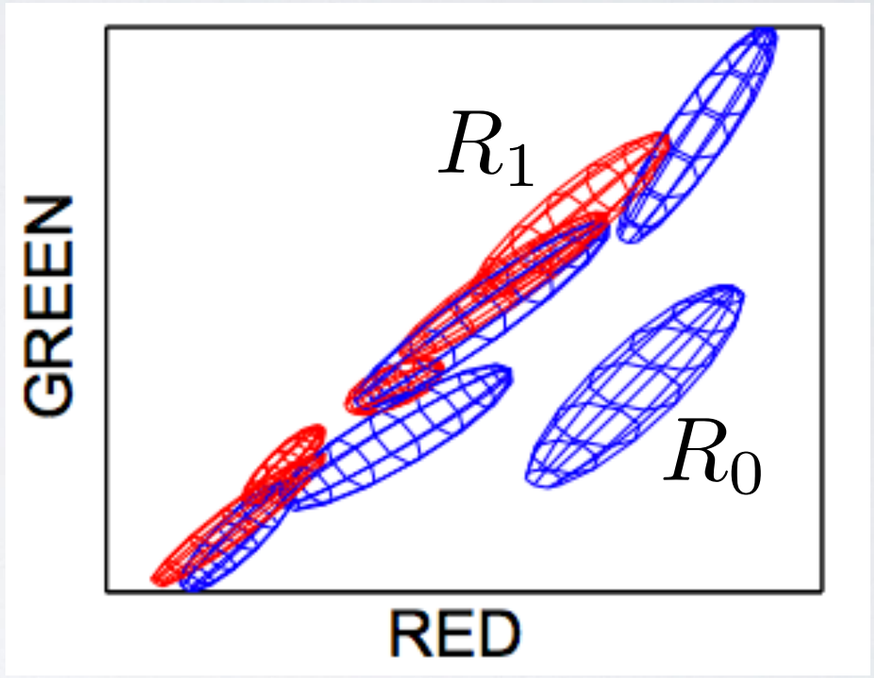




HISTOGRAM MODELS

user marks up image

build gaussian mixture model of regions



$$f(i) = \log p(v_i \in R_0) - \log p(v_i \in R_1)$$

$$\underset{x}{\text{minimize}} \sum_{i,j} |x_i - x_j| w_{i,j} + \sum_i x_i f(i)$$

Rother et. al, "GrabCut," '04

SAMPLE RESULTS

GrabCut (SIGGRAPH '04):

- iterative segmentation & GMM update
- faded alpha channel at edge
- anti-aliasing

Rother et. al, "GrabCut," '04

NUMERICS

graph cuts solves this

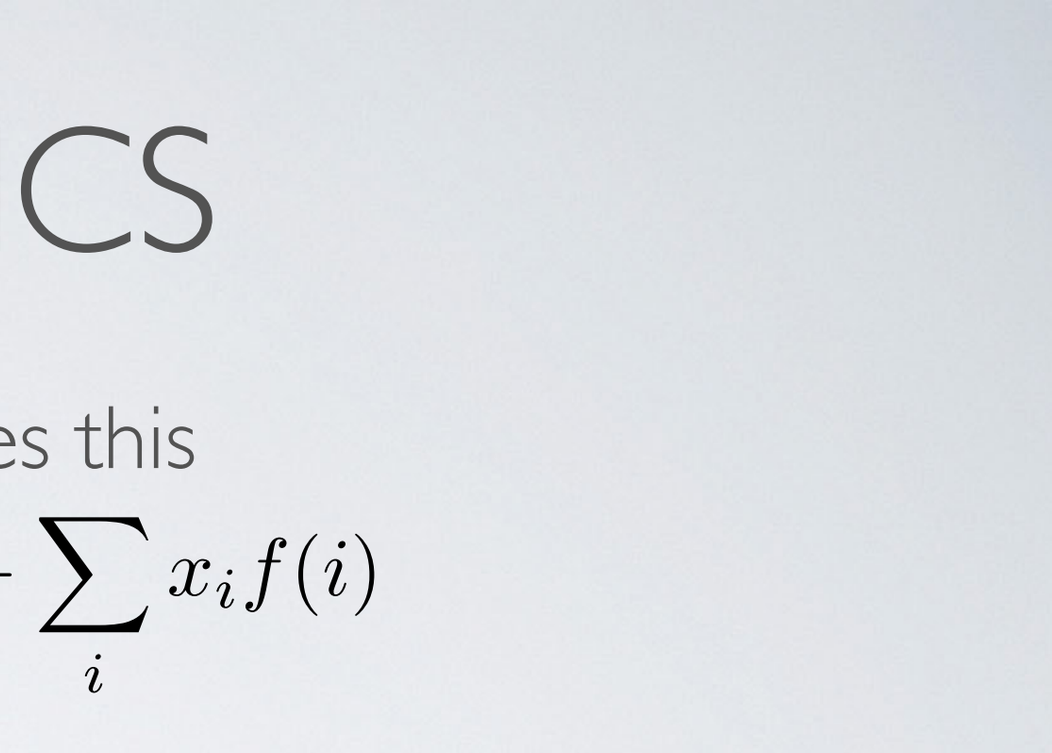
$$\underset{x}{\text{minimize}} \sum_{i,j} |x_i - x_j| w_{i,j} + \sum_i x_i f(i)$$

$$\text{subject to} \;\; x_s = 1, x_t = 0$$

classical methods: **dual** LP

dual(min cut) = max flow

drawback:
hard to implement
does not parallelize well
(bad for GPU)



PDHG APPROACH

graph cuts solves this

$$\underset{x}{\text{minimize}} \sum_{i,j} |x_i - x_j| w_{i,j} + \sum_i x_i f(i)$$

saddle-point form

$$\max_{-w_{ij} \le \lambda_{ij} \le w_{ij}} \min_{0 \le x_{ij} \le 1} \sum_{i,j} \lambda_{ij}(x_i - x_j) + \sum_i x_i f(i)$$

use graph gradient

$$\max_{-w_{ij} \le \lambda_{ij} \le w_{ij}} \min_{0 \le x_{ij} \le 1} \langle \lambda, \nabla x \rangle + \langle x, f \rangle$$

PDHG APPROACH

use graph gradient

$$\max_{-w_{ij} \le \lambda_{ij} \le w_{ij}} \min_{0 \le x_{ij} \le 1} \langle \lambda, \nabla x \rangle + \langle x, f \rangle$$

minimize for x

$$\hat{x} = x^k - \tau(\nabla_g^T \lambda^k + f)$$

$$x^{k+1} = \min\{\max\{\hat{x}, 0\}, 1\}$$

predict

$$\bar{x} = x^{k+1} + (x^{k+1} - x^k)$$

maximize for y

$$\hat{\lambda} = \lambda^k + \sigma \nabla_g \bar{x}$$

$$\lambda_{ij}^{k+1} = \min\{\max\{\hat{\lambda}_{ij}, -w_{ij}\}, w_{ij}\}$$

All entries of primal/dual variables updated simultaneously (GPU)

WHY LINEARIZED PDHG?

ADMM

$$L_\tau(x, \lambda) = f(x) + \langle \lambda, Ax + b \rangle + \frac{\tau}{2} \|Ax + b\|^2$$

must minimize this

PDHG

$$\min_{x \in X} \max_{y \in Y} f(x) + \langle Ax, y \rangle - g(y)$$

evaluate prox

What if you can't do this?

note: linearization applies to all methods discussed above

MODIFIED PROX STEP

$$f(x) + \frac{1}{2\tau}\|x - \hat{x}\|^2$$

replace this distance metric

recall…

$$f(x) \leq f(x^k) + \langle x - x^k, \nabla f(x^k)\rangle + \frac{1}{2\tau}\|x - x^k\|^2$$

small enough

Valid distance metric

$$d(x, x^k) = f(x^k) - f(x) + \langle x - x^k, \nabla f(x^k)\rangle + \frac{1}{2\tau}\|x - x^k\|^2$$

MODIFIED PROX STEP

$$f(x) + \frac{1}{2\tau}\|x - \hat{x}\|^2$$

new prox step

$$x^{k+1} = \min f(x) + d(x, x^k)$$

$$d(x, x^k) = f(x^k) - f(x) + \langle x - x^k, \nabla f(x^k)\rangle + \frac{1}{2\tau}\|x - x^k\|^2$$

simplify

$$x^{k+1} = \arg\min f(x^k) + \langle x - x^k, \nabla f(x^k)\rangle + \frac{1}{2\tau}\|x - x^k\|^2$$

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

LINEARIZED PDHG

$$\max_y \min_x h(x) + f(x) + \langle Ax, y\rangle - g(y)$$

Gradient Descent $$\hat{x} = x^k - \tau(\nabla f(x^k) + A^T y^k)$$

Primal Proximal $$x^{k+1} = \text{prox}_h(\hat{x}, \tau)$$

Predict $$\bar{x} = x_{k+1} + (x_{k+1} - x_k)$$

Gradient Ascent $$\hat{y} = y_k + \tau A\bar{x}$$

Dual Proximal $$y_{k+1} = \text{argmin}\, g(y) + \frac{1}{2\sigma}\|y - \hat{y}\|^2$$

CONVERGENCE THEORY

$$\max_y \min_x h(x) + f(x) + \langle Ax, y \rangle - g(y)$$

Algorithm in dense form

$$x^{k+1} = \text{prox}_h(x^k - \tau(\nabla f(x^k) + A^T y^k))$$

$$y^{k+1} = \text{prox}_g(y^k + \sigma A(2x^{k+1} - x^k))$$

Theorem

The iterates of linearized PDHG converge if the stepsizes satisfy the stability condition

$$\frac{1}{\tau} - \sigma \|A^T A\|_{op} \geq \frac{L_f}{2}.$$

Condat, '13