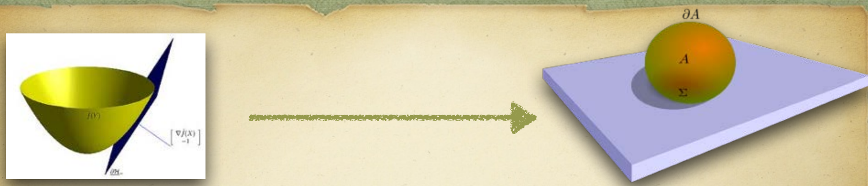


Rethinking Machine Learning in the 21st Century: From Optimization to Equilibration

Sridhar Mahadevan
Autonomous Learning Laboratory
School of Computer Science
University of Massachusetts, Amherst

Part I: Motivation

Autonomous Learning Laboratory

(formerly Adaptive Networks Laboratory, pre 2001)

(Total: 29 graduated PhD students, 5 postdocs, 4 MS students, 3 undergrads)

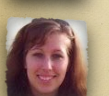
Barto


Lab Directors

Current PhD Students



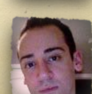








Thomas Boucher
CJ Carey
Bruno Castro da Silva
William Dabney
Stefan Dernbach
Kimberly Ferguson
Ian Gemp
Stephen Giguere
Thomas Helmuth
Nicholas Jacek
Bo Liu
Clemens Rosenbaum
Andrew Stout
Philip Thomas

Notable ALL/ANW Alumni

(Total: 29 graduated PhD students, 5 postdocs, 4 MS students, 3 undergrads)

Michael Jordan (postdoc)
Professor of Statistics, U.C. Berkeley
(h-index: 112)

Richard Sutton
Professor, U. Alberta
(h-index: 53)

Satinder Singh
Professor, U. Michigan
(h-index: 48)

Doina Precup
Associate Professor, McGill
(h-index: 28)

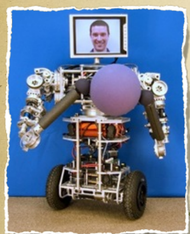
Mohammad Ghavamzadeh
Adobe Research
(h-index: 16)

George Konidaris
Postdoc, MIT
(h-index: 16)

Suchi Saria
Assistant Professor, Johns Hopkins
Mount Holyoke
(PhD: Stanford
(Advisor: Daphne Koller))

Autonomous Learning Lab

LPR Ubot

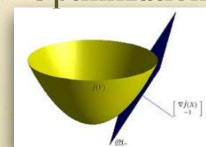


Reinforcement Learning

Curiosity Mars Rover

Supervised Learning

Optimization



Activity recognition

Transfer Learning

Recommendation systems

NETFLIX

Unsupervised Learning

Optimization Problems

Game Theory

Engineering

Linear Programming

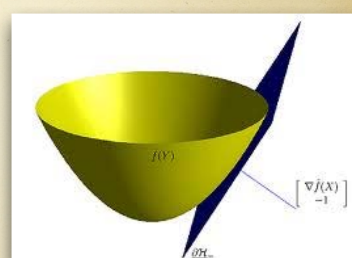
$$\min_x c^T x$$
$$\text{s.t. } Ax = b, x \geq 0$$

ML

Business

Convex Optimization

$x^* = \mathrm{argmin}_x f(x)$ such that $x \in \mathcal{K}$



- If f is a convex function, x^* is its unique minimum whenever

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle, \ \forall x \in K$$

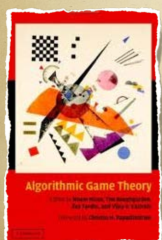
Limitations of Optimization

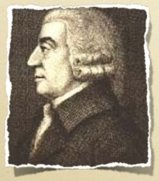
- Single (convex) objective function may not exist
  - World is non-stationary, and competitive
- "Symmetrization" is artificially imposed
  - Similarity matrix in manifold learning
  - Jacobian matrix in gradient optimization

The "Invisible Hand" of the Internet

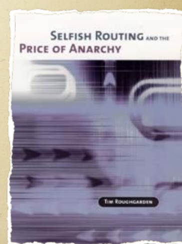
"The Internet is an equilibrium — we just have to identify the game (Scott Shenker)"

"The Internet was the first computational artifact that was not created by a single entity, but emerged from the strategic interaction of many (Christos Papadimitriou)"





Adam Smith
The Wealth of Nations
1776



Changes at IBM

Brendan Mcdermid/Reuters

Virginia M. Rometty, IBM's chief executive, last week announced the company's new Watson division, which will have 2,500 employees.

"This is a key growth area for IBM," said Erich Clementi, senior vice president of IBM Global Technology Services. "We are building out a global footprint." In addition to selling raw computing and data storage capabilities, he said, IBM plans to offer over 150 software and software development products in its cloud. Among the products is Watson, an advanced cognitive computing framework. Last week, IBM's chief executive, Virginia Rometty, announced a new business group inside IBM for Watson.

IBM Plans Big Spending for the Cloud By QUENTIN HARDY JANUARY 16, 2014, NY Times

IBM is moving rapidly on its plans to spend heavily on cloud computing. It expects to spend $1.2 billion this year on increasing the number and quality of computing centers it has worldwide.

The move reflects the speed at which the business of renting a lot of computing power via the Internet is replacing the conventional business of selling mainframe computers, computer servers, and associated hardware and software. Champions of cloud computing cite both lower costs and faster deployment as the reasons for the shift.

MGHPCC
MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER

http://www.mghpcc.org/

http://www.bu.edu/

http://www.harvard.edu/

http://web.mit.edu/

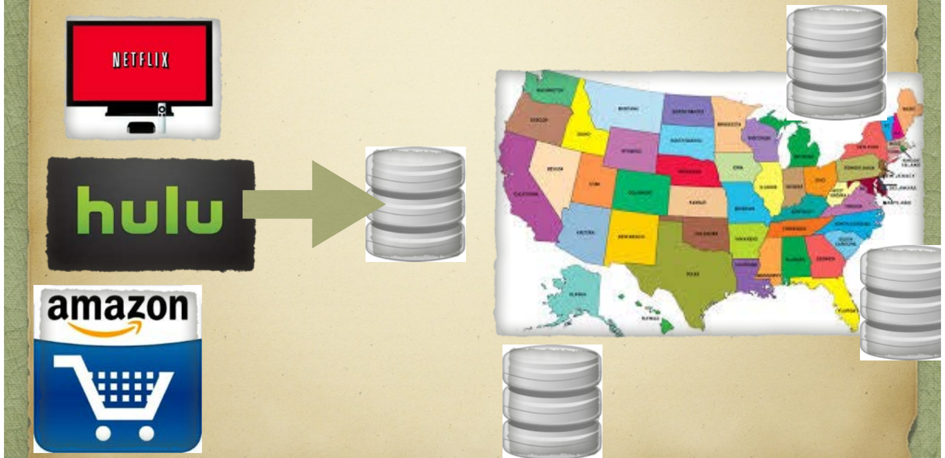
http://www.massachusetts.edu/
http://www.northeastern.edu/

http://www.cisco.com/

http://www.emc.com/

"Netflix" Cache Problem

(Dernbach, Kurose, Mahadevan, Technicolor)



Next Generation Internet Model [Nagurney et al., 2014]

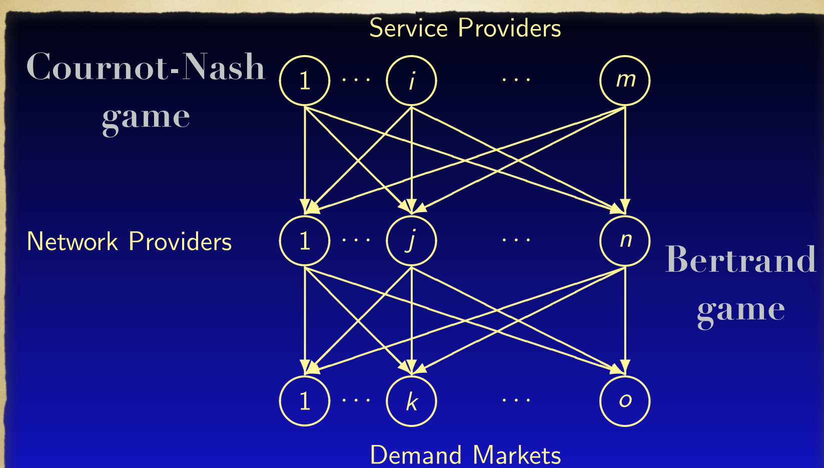
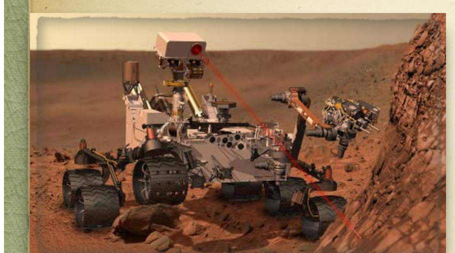


Figure 1: The Network Structure of the Cournot-Nash-Bertrand Model for a Service-Oriented Internet

Multiple data sources on Mars Curiosity Rover



Rock Abrasion Tool

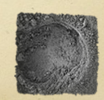
Miniature Thermal Emission Spectrometer

Moessbauer Spectrometer

Alpha Particle X-ray Spectrometer

Microscopic Imager

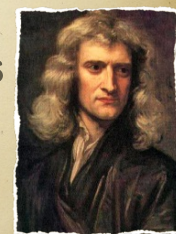
How to handle competition across across instruments and scientists?



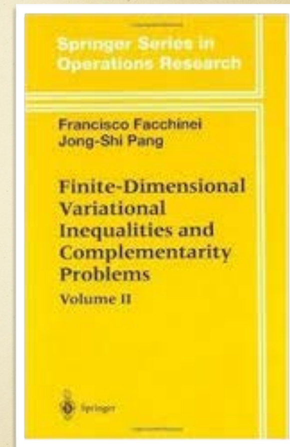
Part II: A New Framework for ML

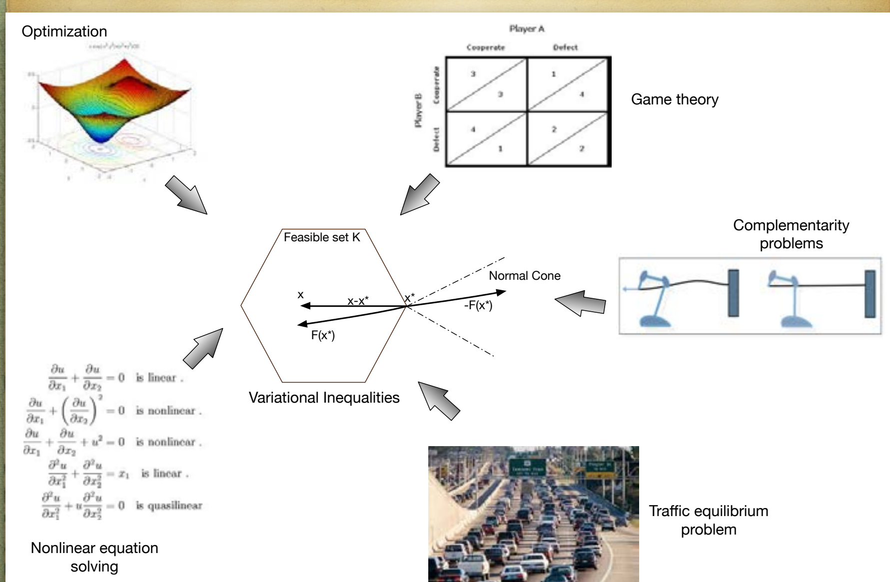
"If I have seen further it is by standing on ye sholders of Giants"

Letter to Robert Hooke (15 February 1676 Isaac Newton)

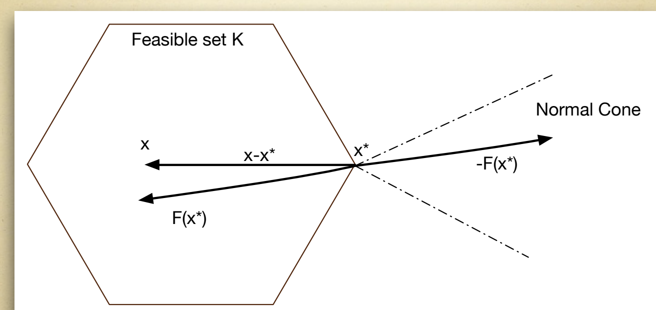


Guido Stampacchia



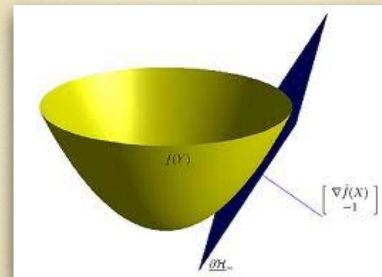


Variational Inequality



$$\langle F(x^*), x - x^* \rangle \geq 0, \ \forall x \in K$$

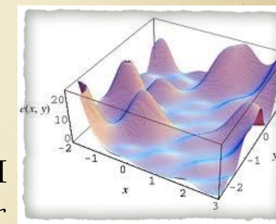
Convex Optimization => VI



$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle, \ \forall x \in K$$

Optimization => VI

Suppose $x^* = \text{argmin}_{x \in \mathcal{K}} f(x)$
where f is differentiable



Then x^* solves the VI
$\langle \nabla f(x^*), x - x^* \rangle \geq 0. \ \forall x \in \mathcal{K}$

Proof: Define $\phi(t) = f(x^* + t(x - x^*))$
Since $\phi(0)$ achieves the minimum
$\phi'(0) = \langle \nabla f(x^*), x - x^* \rangle \geq 0$

When VI => optimization?

Given $VI(F,K)$, define $\nabla F(x) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial F_n}{\partial x_1} & \cdots & \frac{\partial F_n}{\partial x_n} \end{bmatrix}$

When ∇F is symmetric and positive semi-definite
VI(F,K) can be reduced to an optimization problem,

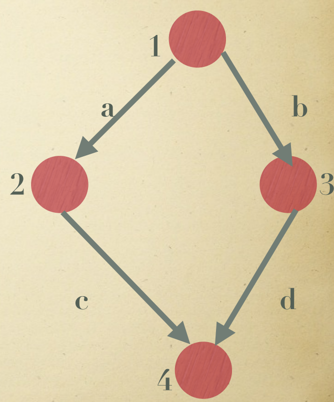
Optimization vs VIs

Property	Optimization	VI
Mapping	(Strong) Convexity	(Strong) Monotonicity
Jacobian	Positive definite and symmetric	Asymmetric
Objective function	Single fixed	Multiple or none

Traffic Network Equilibrium

(Dafermos, Nagurney)

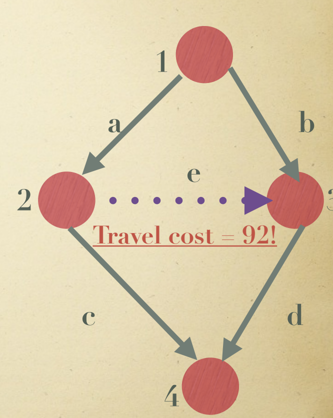
- Link travel cost functions
  - $C_a(F_a) = 10*F_a$
  - $C_b(F_b) = F_b + 50$
  - $C_c(F_c) = F_c + 50$
  - $C_d(F_d) = 10\ F_d$
- Travel demand $D_{14} = 6$
- Find equilibrium flows



Traffic Network Equilibrium

(Dafermos, Nagurney)

- Flows at equilibrium
  - $F_a = F_b = 3$
  - $F_c = F_d = 3$
  - $C_a = 30$, $C_b = 53$
  - $C_c = 53$, $C_d = 30$
- Path costs = 83
- Nash equilibrium



Part II: Algorithms

Composite Objective Functions from recent ALL Research

"Sparse" Supervised learning

Lasso: $\min_{x \in X} f(x) + g(x) : \min_{\beta \in \mathbb{R}^k} \|X\beta - y\|_2^2 + \lambda \|\beta\|_1$

RO-TD: "Saddle Point" Reinforcement Learning

$$\min_x \|Ax - b\|_m + h(x) = \min_x \max_{\|y\|_n \leq 1} y^T(Ax - b) + h(x)$$

Unsupervised learning

Low-rank embedding: $\min_R \frac{1}{2} ||X - XR||_F^2 + \lambda ||R||_*$

$0 \in T(x)$
Monotone Inclusion

$0 \in A(x) + B(x)$
Operator Splitting

$\mathrm{prox}_f(v) = \mathrm{argmin}_x(f(x) + \frac{1}{2}\|x - v\|_2^2)$
Proximal Mappings

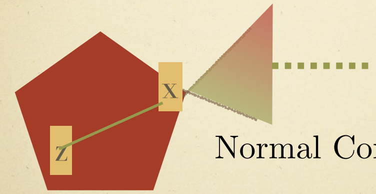
$x_{k+1} \leftarrow \nabla\psi^*(\nabla\psi(x_k) - \alpha_k \partial f(x))$
Mirror Descent

ADMM
(Dual Decomposition)

Normal Cone

Subdifferential of a convex function:

$$\partial f(x) = \{v \in \mathbb{R}^n : f(z) \geq f(x) + v^T(z - x), \forall z \in \mathrm{dom}(f)\}$$

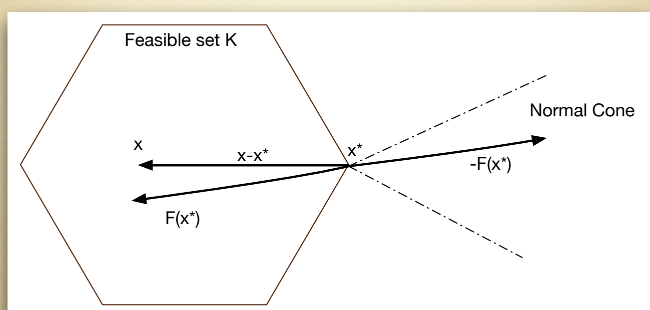


Normal Cone: $\partial I_C(x) = N_C(x)$

$I_C(x)$

VI as monotone inclusion

$$0 \in F(x^*) + N_K(x^*)$$

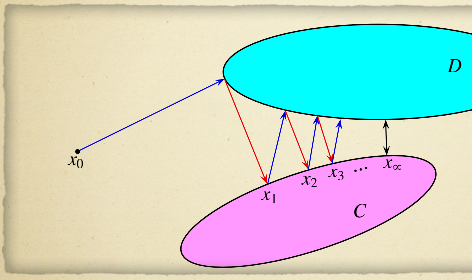


Distributed Optimization via ADMM

(Boyd et al., ML FT 2010)

"ADMM was developed over a generation ago, with its roots stretching far in advance of the Internet, distributed and cloud computing systems, massive high-dimensional datasets, and associated large-scale applied statistical problems. Despite this, it appears well-suited to the modern regime."

Convex Feasibility Problem

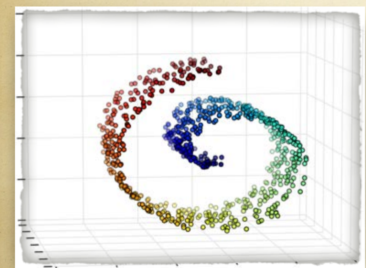


Operator Splitting

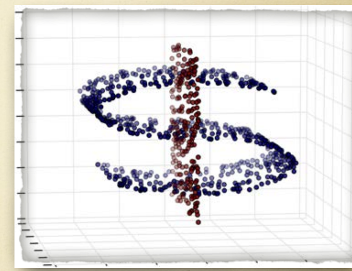
$$x_{k+1} \leftarrow (I + \lambda A)^{-1}(I - \lambda B)x_k$$

Proximal splitting methods in signal processing
Combetti and Pesquet

Manifold Learning



Single Manifold
(LLE, ISOMAP, Diffusion Maps,
Laplacian Eigenmaps)



Mixture of Manifolds
(Low-rank embedding)

Manifold Warping

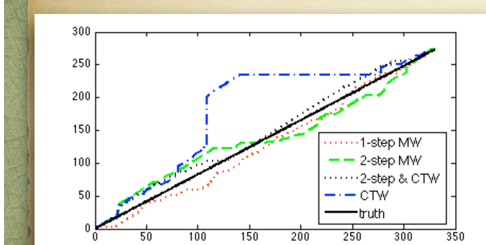
(Vu, Carey, and Mahadevan, AAAI 2012)

- Combine dynamic time warping and manifold alignment using alternating projections
- Minimize the loss function to preserve local geometry and correspondences

$$\begin{aligned} L_1\left(F^{(X)}, F^{(Y)}\right) &= \mu \sum_{i \in X, j \in Y} ||F_i^{(X)} - F_j^{(Y)}||^2 W_{i,j}^{(X,Y)} \\ &+ (1-\mu) \sum_{i,j \in X} ||F_i^{(X)} - F_j^{(X)}||^2 W_{i,j}^{(X)} \\ &+ (1-\mu) \sum_{i,j \in Y} ||F_i^{(Y)} - F_j^{(Y)}||^2 W_{i,j}^{(Y)} \end{aligned}$$

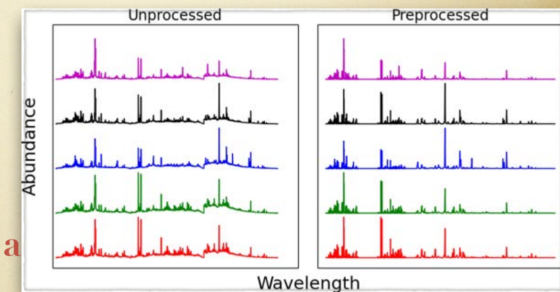
Manifold Alignment over time

- CMU Multimodal activity dataset
- Measure human activity while cooking
- 26 subjects
- 5 different recipes



MARS Curiosity Rover

Mineral Spectra



Curiosity zapping a rock with a laser

Boucher, Carey, Darby, Mahadevan, 2014

Low-Rank Alignment

Step 1: Compute Reconstructions

$$\min_R \frac{1}{2}||X - XR||_F^2 + \lambda||R||_*$$

(ADMM)

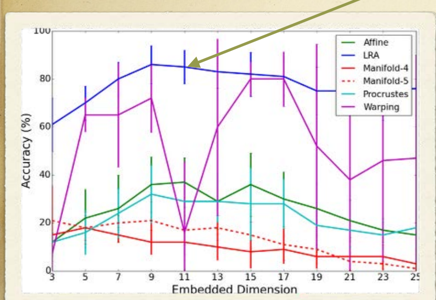
Step 2: Compute Low-Dimensional Embeddings

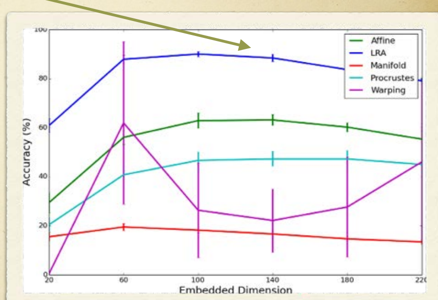
$$\min_{F^{(X)}} \frac{1}{2}||F^{(X)} - F^{(X)}R||_F^2 \ s.t. \ (F^{(X)})^\top F^{(X)} = I.$$

Eigen Decomposition

Experimental Results

Astronomy LRA Cross-lingual Alignment





Martian Spectroscopy

EU Parallel Corpus English-German

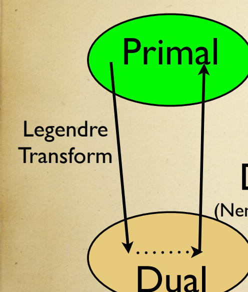
Boucher, Carey, Darby, Mahadevan, 2014

Optimization in High-Dimensions

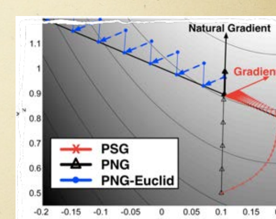
[Thomas, Dabney, Mahadevan, Giguerre, NIPS 2013]

$\text{argmin}_{x \in X} f(x)$

Natural Gradient Descent (Amari)



Mirror Descent
(Nemirovski and Yudin)



$x_{k+1} \leftarrow \nabla\psi_k^*(\nabla\psi(x_k) - \alpha_k \partial f(x_k))$

$x_{k+1} \leftarrow x_k - \alpha_k G_k^{-1} \nabla f(x_k)$

Mirror Descent = Natural Gradient!

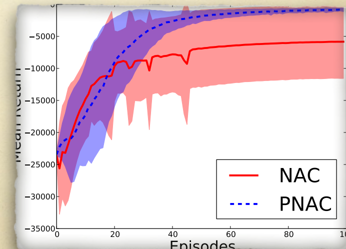
[Thomas, Dabney, Mahadevan, Giguerre, NIPS 2013]

Theorem 5.1. *The natural gradient descent update at step k with metric tensor $G_k \triangleq G(x_k)$:*

$$x_{k+1} = x_k - \alpha_k G_k^{-1} \nabla f(x_k),$$

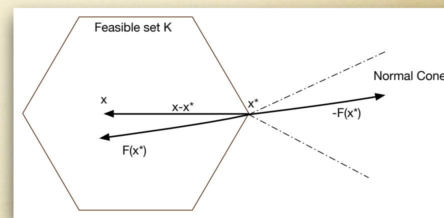
is equivalent to (1)*, the mirror descent update at step k, with $\psi_k(x) = (1/2)x^\intercal G_k x$.*

$$x_{k+1} = \nabla \psi_k^*(\nabla \psi_k(x_k) - \alpha_k \nabla f(x_k)), \qquad (1)$$



Fixed Point Formulation

- Let Π_K be the projection onto convex set K
- Then x* solves VI(F,K) if and only if x* is the fixed point of the mapping given by
- $x^* = \Pi_K(x^* - \gamma F(x^*))$



Projection Algorithm

Algorithm 1 The Basic Projection Algorithm for solving VIs.

INPUT: Given VI(F,K), and a symmetric positive definite matrix D.

1: Set $k = 0$ and $x_k \in K$.
2: **repeat**
3: Set $x_{k+1} \leftarrow \Pi_{K,D}(x_k - D^{-1}F(x_k))$.
4: Set $k \leftarrow k + 1$.
5: **until** $x_k = \Pi_{K,D}(x_k - D^{-1}F(x_k))$.
6: Return x_k

Monotonicity Properties

Strongly monotone mapping:

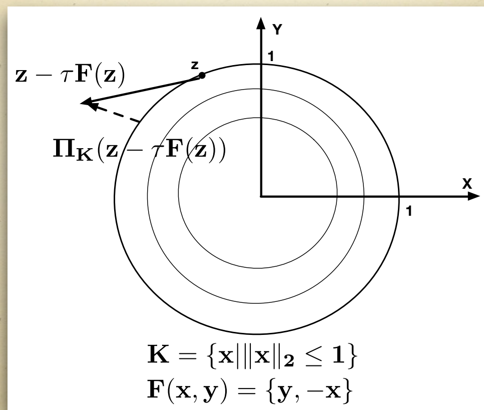
$$\langle F(x) - F(y), x - y \rangle \geq \mu \|x - y\|_2^2, \mu > 0, \forall x, y \in K$$

Lipschitz mapping:

$$\|F(x) - F(y)\|_2 \leq L\|x - y\|_2, \forall x, y \in K$$

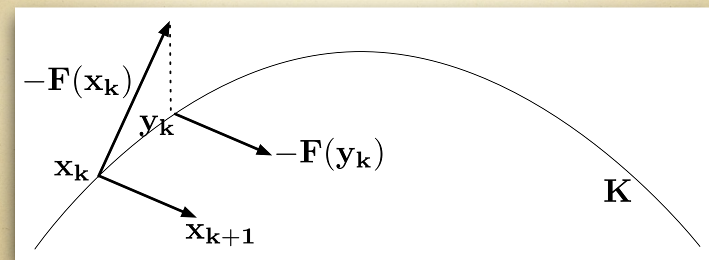
Example: if F is the gradient map of a function f, then strong monotonicity of F implies f is strongly convex

Projection Method Fails



Bertsekas and Tsitsiklis, Parallel and Distributed Computation, Athena Scientific.

Extragradient Method



Korpolevich developed the extragradient method, which is the most popular method for solving VIs

Extragradient Method

Algorithm 2 The Extragradient Algorithm for solving VIs.

INPUT: Given VI(F,K), and a scalar α.

Set $k = 0$ and $x_k \in K$.
repeat
  Set $y_k \leftarrow \Pi_K(x_k - \alpha F(x_k))$.
  Set $x_{k+1} \leftarrow \Pi_K(x_k - \alpha F(y_k))$.
  Set $k \leftarrow k + 1$.
until $x_k = \Pi_K(x_k - \alpha F(x_k))$.
Return x_k

Khobotov developed a learning rate rule under which the extragradient method works for all pseudo-monotone mappings

$$\langle F(y), x - y\rangle \geq 0 \Rightarrow \langle F(x), x - y\rangle \geq 0, \ \forall x, y \in K$$

Runge-Kutta Method for VIs (Ian Gemp)

Runge Kutta (4) Gradient Descent

$k_1 = \alpha\nabla F(x_k)$
$k_2 = \alpha\nabla F(x_k - \frac{1}{2}k_1)$
$k_3 = \alpha\nabla F(x_k - \frac{1}{2}k_2)$
$k_4 = \alpha\nabla F(x_k - k_3)$

$x_{k+1} = x_k - \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$

Runge Kutta (4) Non-Euclidean Extragradient

$k_1 = \alpha F(x_k)$
$k_2 = \alpha F(\nabla\psi_k^*(\nabla\psi_k(x_k) - \frac{\alpha}{2}k_1))$
$k_3 = \alpha F(\nabla\psi_k^*(\nabla\psi_k(x_k) - \frac{\alpha}{2}k_2))$
$k_4 = \alpha F(\nabla\psi_k^*(\nabla\psi_k(x_k) - \alpha k_3))$
$x_{k+1} = \nabla\psi_k^*(\nabla\psi_k(x_k) - \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4))$

General Runge Kutta Gradient Descent

$k_1 = \alpha\nabla F(x_k)$
$k_2 = \alpha\nabla F(x_k - a_{21}k_1)$
$k_3 = \alpha\nabla F(x_k - a_{31}k_1 - a_{32}k_2)$
$\vdots$
$k_s = \alpha\nabla F(x_k - a_{s1}k_1 - a_{s2}k_2 - \ldots - a_{s,s-1}k_{s-1})$

$x_{k+1} = x_k - \Sigma_{i=1}^s b_i k_i$

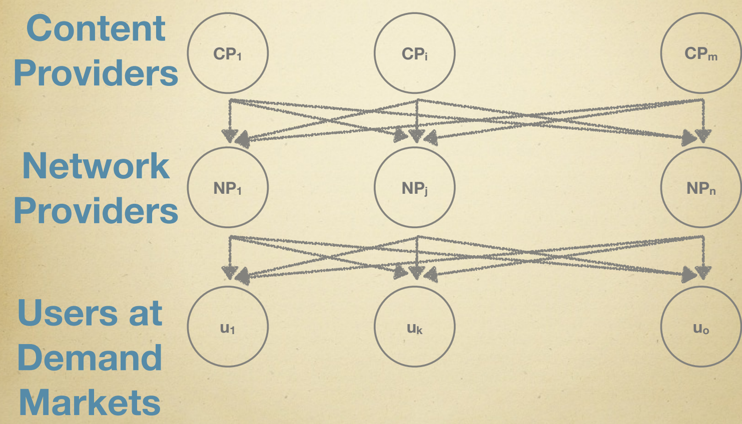
General RK Non-Euclidean Extragradient

$k_1 = \alpha F(x_k)$
$k_2 = \alpha F(\nabla\psi_k^*(\nabla\psi_k(x_k) - a_{21}k_1))$
$k_3 = \alpha F(\nabla\psi_k^*(\nabla\psi_k(x_k) - a_{31}k_1 - a_{32}k_2))$
$\vdots$
$k_s = \alpha F(\nabla\psi_k^*(\nabla\psi_k(x_k) - a_{s1}k_1 - a_{s2}k_2 - \ldots - a_{s,s-1}k_{s-1}))$

$x_{k+1} = \nabla\psi_k^*(\nabla\psi_k(x_k) - \Sigma_{i=1}^s b_i k_i)$

Next-Generation Internet

(Nagurney et al., 2014)



VI Formulation

- Production cost function *f*(Q) - cost of providing a certain volume of content
- Demand price function ρ(Q,q) - user offer depends on content quality and market volume

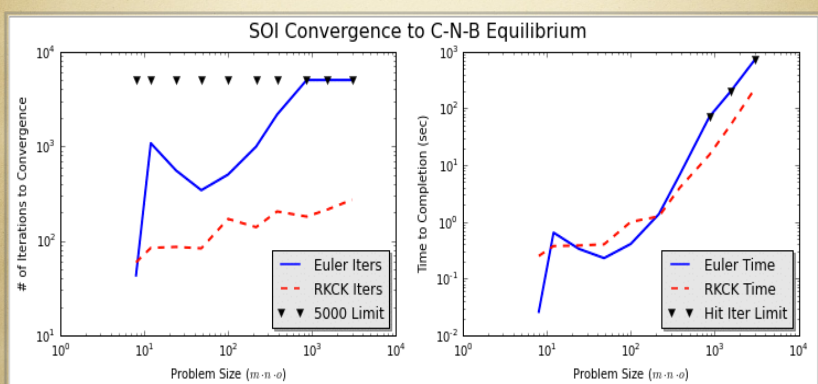
$$\langle F(X^*), X - X^* \rangle \geq 0, \quad \forall X \in \mathcal{K}, \qquad X \equiv (Q, q, \pi)$$

$$F^1_{ijk}(X) = \frac{\partial \hat{f}_i(Q)}{\partial Q_{ijk}} + \pi_{ijk} - \hat{\rho}_{ijk}(Q,q) - \sum_{h=1}^{n}\sum_{l=1}^{o} \frac{\partial \hat{\rho}_{ihl}(Q,q)}{\partial Q_{ijk}} \times Q_{ihl},$$

$$F^2_{ijk}(X) = \sum_{h=1}^{m}\sum_{l=1}^{o} \frac{\partial c_{hjl}(Q,q)}{\partial q_{ijk}},$$

$$F^3_{ijk}(X) = -Q_{ijk} + \frac{\partial oc_{ijk}(\pi_{ijk})}{\partial \pi_{ijk}}.$$

Results of Runge-Kutta on Internet VI Problem



Projected Dynamical Systems

Two Player Game

Row	Column		
		Heads	Tails
	Heads	(1, −1)	(−1, 1)
	Tails	(−1, 1)	(1, −1)

Classical Dynamical System

$$\alpha_{t+1} = \alpha_t + \eta_t \frac{\partial V_r(\alpha_t, \beta_t)}{\partial \alpha}$$

$$\beta_{t+1} = \beta_t + \eta_t \frac{\partial V_c(\alpha_t, \beta_t)}{\partial \beta}$$

Projected Dynamical System

$$\dot{x} = \Pi_K(x, -F(x)), x(0) = x_0 \in K$$

Projected dynamical systems are a more powerful framework for studying dynamics of equilibria in games than classical dynamical systems used in [Singh et al., UAI 2000]

PDS Formulation

ODE/IVP

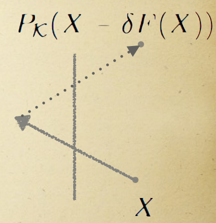
$$\dot{X} = \Pi_{\mathcal{K}}(X, -F(X)), \quad X(0) = X^0,$$

Projection Operator

$$\Pi_{\mathcal{K}}(X, -F(X)) = \lim_{\delta \to 0} \frac{P_{\mathcal{K}}(X - \delta F(X)) - X}{\delta}$$

Fixed Point Problem

$$\dot{X} = 0 = \Pi_{\mathcal{K}}(X^*, -F(X^*))$$

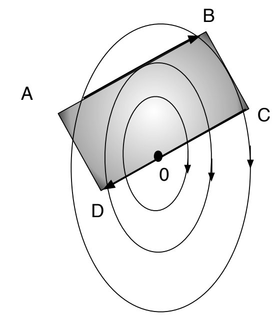


- X* solves the VI iff it is a stationary point of the projected ODE
- Lipschitz continuity of F(X) guarantees the existence of a unique solution
- Stability of equilibrium is given by the monotonicity of F(X) which can be determined from the positive-definiteness of the Jacobian of F(X)

Skorokhod Analysis

$$\dot{\phi}_x(t) = \Pi_K(\phi_x(t), -F(\phi_x(t))), \;\; \phi_x(0) = x$$

$$F_1(x_1, x_2) = -x_2, F_2(x_1, x_2) = 4x_1$$



Alternating Direction Method of Multipliers

Minimize $f(x) + g(x)$

Solve $0 \in \partial f(x) + \partial g(x)$

Choose $A(x) = \partial g(x), B(x) = \partial f(x)$

ADMM is an instance of Douglas Rachford splitting

$$x_{k+\frac{1}{2}} = \operatorname{argmin}_x (f(x) + \frac{1}{2\lambda}\|x - z_k\|_2^2)$$

$$z_{k+\frac{1}{2}} = 2x_{k+\frac{1}{2}} - z_k$$

$$x_{k+1} = \operatorname{argmin}_x (g(x) + \frac{1}{2\lambda}\|x - x_{k+\frac{1}{2}}\|_2^2)$$

$$z_{k+1} = z_k + x_{k+1} - x_{k+\frac{1}{2}}$$

ADMM for Cloud Computing

Algorithm 2 An iteration of global consensus ADMM in Hadoop/ MapReduce.

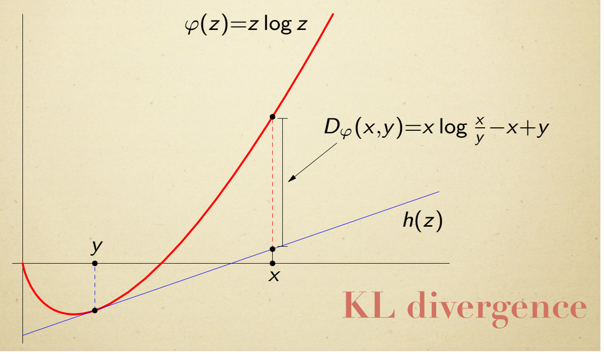
function map(key i, dataset $\mathcal{D}_i$)
1. Read $(x_i, u_i, \hat{z})$ from HBase table.
2. Compute $z := \mathbf{prox}_{g,N\rho}((1/N)\hat{z})$.
3. Update $u_i := u_i + x_i - z$.
4. Update $x_i := \operatorname{argmin}_x \left(f_i(x) + (\rho/2)\|x - z + u_i\|_2^2\right)$.
5. *Emit* (key CENTRAL, record (x_i, u_i)).

function reduce(key CENTRAL, records $(x_1, u_1), \ldots, (x_N, u_N)$)
1. Update $\hat{z} := \sum_{i=1}^N x_i + u_i$.
2. *Emit* (key j, record $(x_j, u_j, \hat{z})$) to HBase for $j = 1, \ldots, N$.

Boyd et al., ML Fn Trends, 2010

Bregman Divergence

$$D_\varphi(\mathbf{x},\mathbf{y}) = \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - (\mathbf{x}-\mathbf{y})^T \nabla\varphi(\mathbf{y})$$



Bregman ADMM

"There is no known proof of convergence known for ADMM with non-quadratic penalty terms", Boyd et al., 2010

Wang and Banerji, 2013:

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \langle \mathbf{y}_t, \mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c}\rangle + \rho B_\phi(\mathbf{c} - \mathbf{Ax}, \mathbf{Bz}_t) + \rho_{\mathbf{x}} B_{\varphi_{\mathbf{x}}}(\mathbf{x}, \mathbf{x}_t)\,,$$
$$\mathbf{z}_{t+1} = \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} g(\mathbf{z}) + \langle \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz} - \mathbf{c}\rangle + \rho B_\phi(\mathbf{Bz}, \mathbf{c} - \mathbf{Ax}_{t+1}) + \rho_{\mathbf{z}} B_{\varphi_{\mathbf{z}}}(\mathbf{z}, \mathbf{z}_t)\,,$$
$$\mathbf{y}_{t+1} = \mathbf{y}_t + \tau(\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c})\,.$$

Bauschke et al., 2004: $\overleftarrow{\operatorname{prox}}_\varphi : y \mapsto \operatorname*{argmin}_{x\in U} \varphi(x) + D(x,y)$

$$\overrightarrow{\operatorname{prox}}_\psi : x \mapsto \operatorname*{argmin}_{y\in U} \psi(y) + D(x,y).$$

fix $x_0 \in U$ and set $(\forall n \in \mathbb{N})\quad y_n = \overrightarrow{\operatorname{prox}}_\psi(x_n)$ and $x_{n+1} = \overleftarrow{\operatorname{prox}}_\varphi(y_n)$.

Generalized ADMM Method for Separable VIs

(Tseng, 1988)

$$\langle x - x^*, R(x^*)\rangle + \langle z - z^*, S(z^*)\rangle \geq 0,$$
$$\forall (x,z) \in X \times Z \; s.t. \; Ax + Bz = b$$

↓

Minimize $\langle R(x^*), x\rangle + \langle S(z^*), z\rangle$

$s.t. \; x \in X, z \in Z, Ax + Bz = b$

↓

Let $N(.|X), N(.|Z)$ be subdifferentials of $\delta(.|X), \delta(.|Z)$

↓

Let p^* be the optimal Lagrange multiplier for $Ax + Bz = b$

Generalized ADMM for Separable VIs

Karush Kuhn Tucker conditions imply:

$$A^T p^* \in N(x^*|X) + R(x^*)$$
$$B^T p^* \in N(z^*|Z) + S(z^*)$$
$$Ax^* + Bz^* = b$$

Define maximal monotone operators

$$F(x) = R(x) + N(x|X)$$
$$G(z) = S(z) + N(z|Z)$$

Above equations can be rewritten as:

$$AF^{-1}(A^T p^*) + BG^{-1}(B^T p^*) = b$$

Splitting Algorithm for Separable VIs

Find x_t $s.t.$ $\langle x - x_t, R(x_t) - A^T p(t) \rangle \geq 0, \forall x \in X$

$$\text{Compute } z_t \ s.t.\ \langle z - z_t, S(z_t) - B^T(p(t) - c(t)(Ax_t + Bz_t - b)) \rangle \geq 0, \forall z \in Z$$

Update $p(t+1) = p(t) + c(t)(b - Ax_t - Bz_t)$

Summary

- VIs and PDS provide a new direction for ML research
- Many applications and challenges
  - Non-cooperative version of distributed ADMM optimization

Questions?

Game theory => VI

- A CN game consists of m players, where player i chooses a strategy $x_i \in X_i$
- Let the joint payoffs for player i be $F_i(x_1,\ldots,x_m)$
- A set of strategies x* is in Nash equilibrium if $\langle (x_i - x_i^*), \nabla_i F_i(x_i^*) \rangle \geq 0$