Imperial College London

Optimization by Machine Learning for Intelligent Communication Networks

Queen's Tower
Imperial College

Kin K. Leung

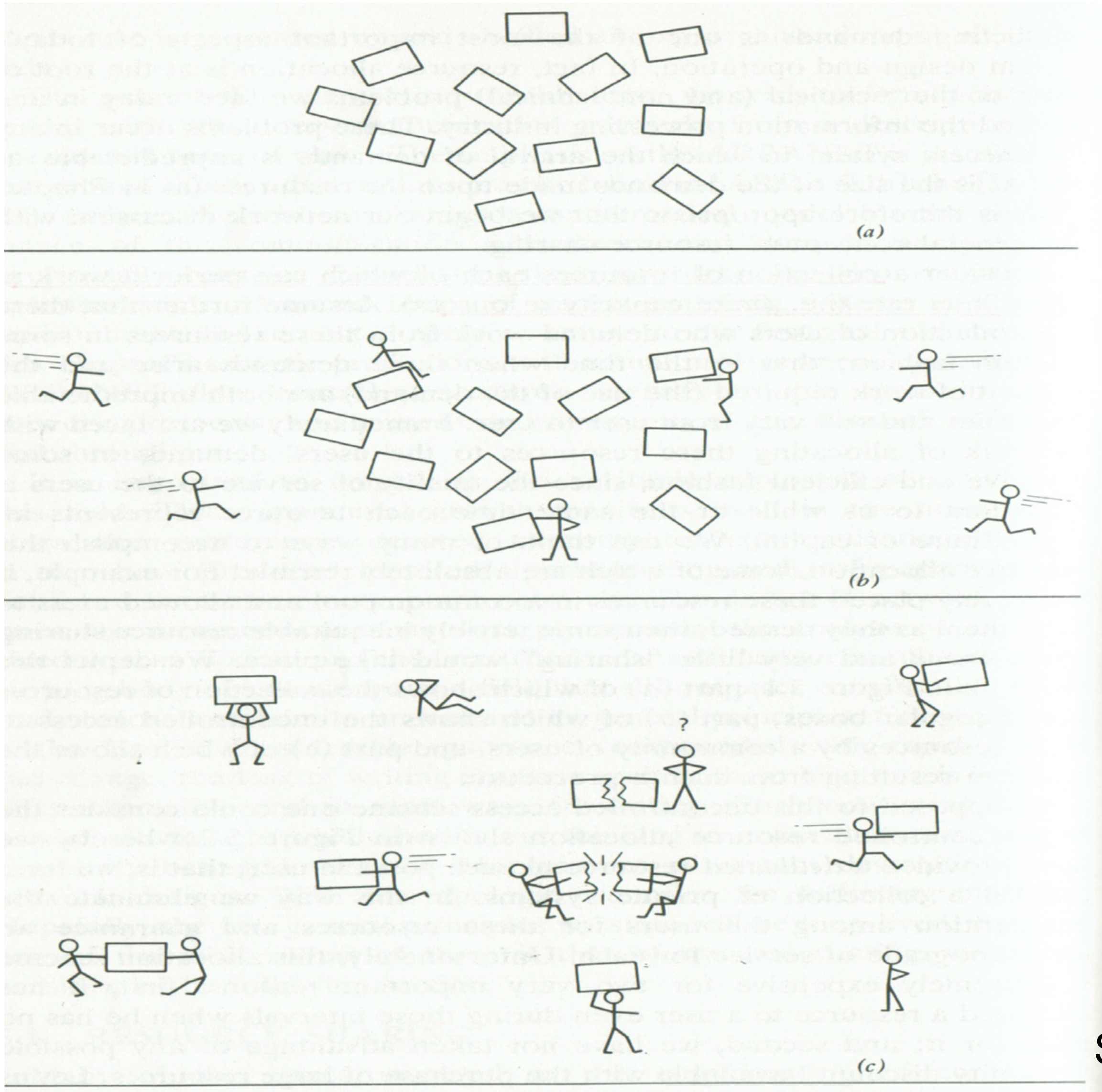
Electrical & Electronic Engineering, and Computing Departments

Imperial College, London

May 2023

Acknowledgments: Zheyu (Joe) Chen, Sepideh Nazemi, Faheem Zarfari, George Tychogiorgos (Imperial College), Ananthram Swami and Kevin Chan (U.S. Army Research Lab), Don Towsley (UMass), Shiqiang Wang (IBM US), Leandros Tassiulas (Yale), Patrick Baker (UK Dstl / Royal Air Force)

Resource Allocation or Sharing



Allocation Criteria:

Capacity/availability

Quality of service

Fairness

Price

…

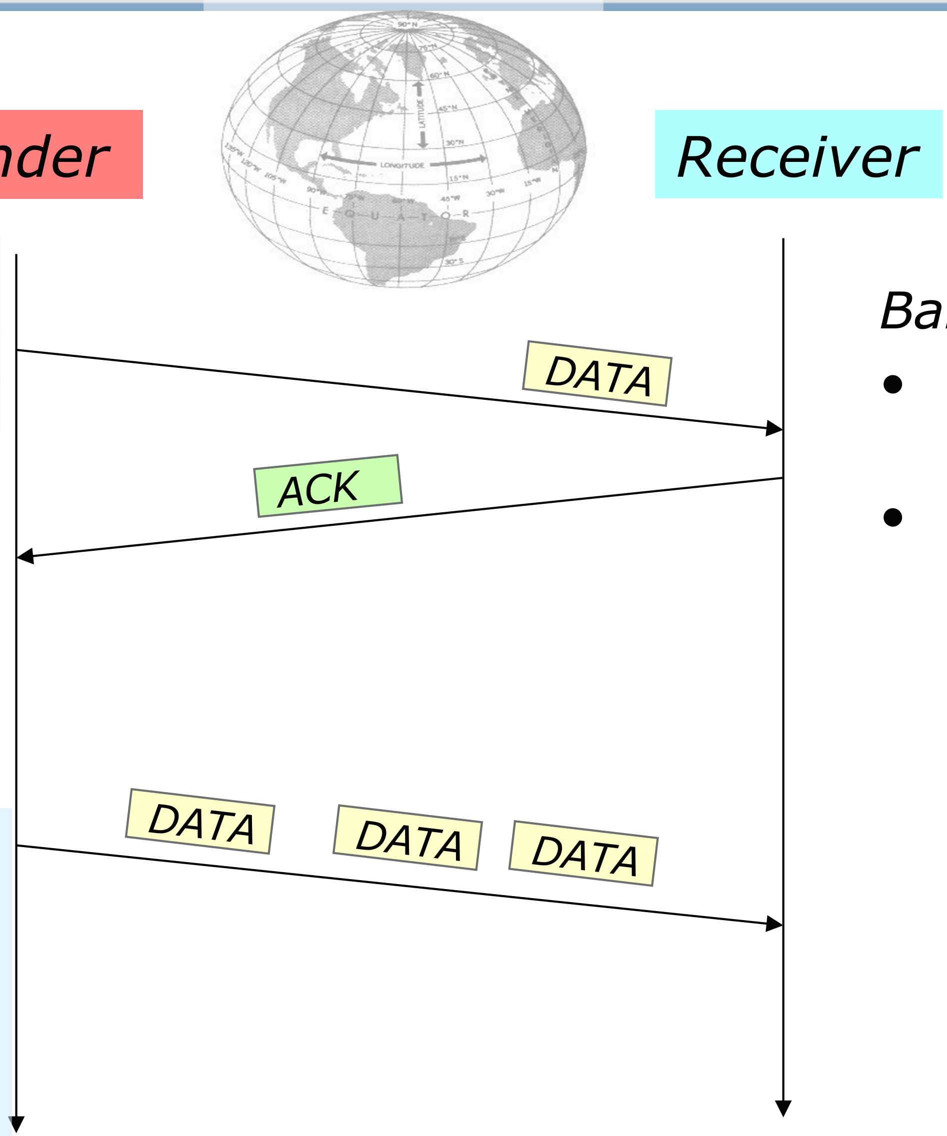
Optimality = ???

=> Mathematical formulation

Source: Kleinrock (1976)

Imperial College London

Transport Control Protocol (TCP): Distributed Resource Allocation



Set timeout when send packet

Track round trip times for future timeout values

Re-send if timeout

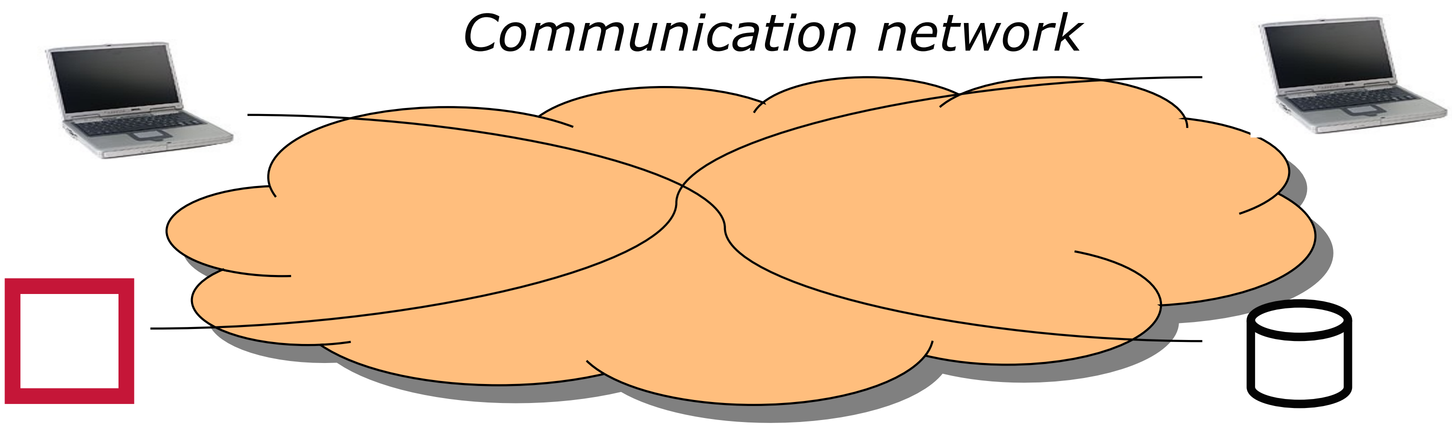
Dynamically adjusted number of segments outstanding at a time (Congestion window); Timeout reduces window size to 1

Bandwidth allocation:

- *Control by congestion window*
- *Window size depends on delay, packet loss, etc.*

TCP = distributed resource allocation algorithm

Distributed Bandwidth Allocation: Network Utility Maximization (NUM)



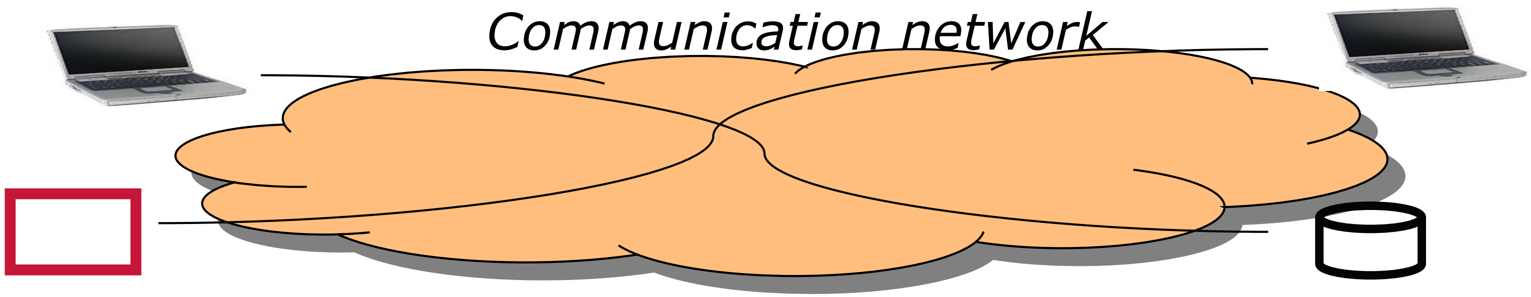
- Allocate resource (bandwidth) as a convex optimization problem

$$\max_{\underline{x}} \sum_i U_i(\underline{x}) \quad \text{subject to} \quad A\underline{x} \leq C \text{ and } \underline{x} \geq 0$$

where each user *i* has a fixed communication path and a utility function $U_i(\underline{x})$ and is allocated with data rate x_i . Bandwidth allocated to all users must be less than link capacity *C*.

See Kelly, Maulloo and Tan (1998)

TCP: Distributed Optimization of Resource Allocation



- TCP allocates resources (bandwidth, buffer, etc.) to optimize

$$\max_{\underline{x}} \sum_i U_i(\underline{x}) \quad \text{subject to} \quad A\underline{x} \leq C \text{ and } \underline{x} \geq 0$$

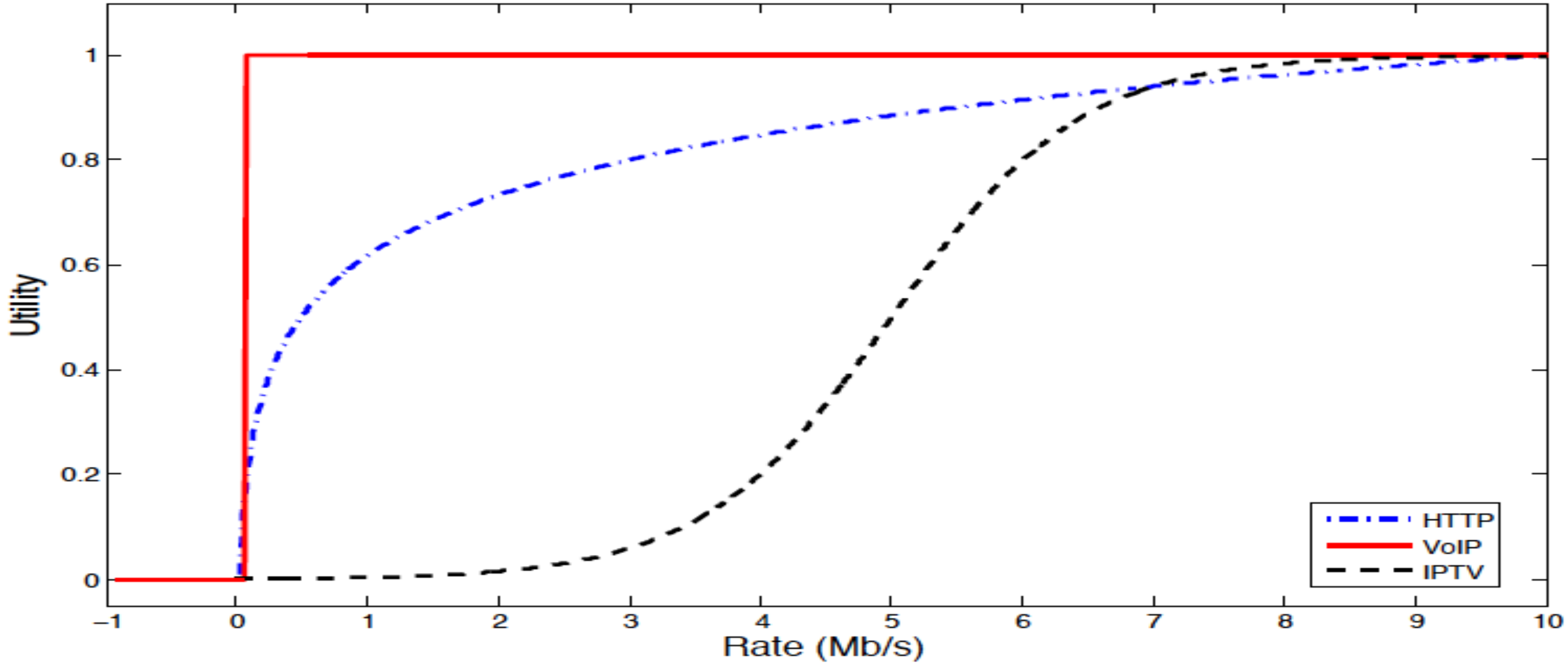
Primal iterations set source rates; dual iterations by active queue management (AQM) protocol

TCP Version	Utility Function
TCP Reno-1	$U(x_s) = \frac{\sqrt{\frac{3}{2}}}{D_s} \tan^{-1}\left(\sqrt{\frac{2}{3}} x_s D_s\right)$
TCP Reno-2	$U(x_s) = \frac{1}{D_s} \log \frac{x_s D_s}{2x_s D_s + 3}$
TCP Vegas	$U(x_s) = \alpha_s d_s \log x_s$

These utility functions are concave for convex optimization!

See Low et al. (2000, 2002, 2003)

Need for non-concave utility functions



Application	Utility Function
HTTP	$U(x_s) = U_{max} \frac{\log\left(\frac{x_s}{x_s^{min}}\right)}{\log\left(\frac{x_s^{max}}{x_s^{min}}\right)} \frac{sgn(x_s - x_s^{min})+1}{2}, \ 0 \le x_s \le x_s^{max}$
VoIP	$U(x_s) = U_{max} \frac{sgn(x_s - x_s^{min})+1}{2}, \ 0 \le x_s$
IPTV	$U(x_s) = \frac{U_{max}}{1+\frac{1}{\epsilon-1}e^{-x_s \cdot \alpha}}, \ \alpha = \frac{2\ln\left(\frac{1}{\epsilon-1}\right)}{x_s^{max}}$ and $0 \le x_s \le x_s^{max}$

Applications may have non-TCP utilities, thus TCP may not allocate resources optimally!

Sufficient Condition for Solving Non-convex Problem

The primal problem:

$$\max_{\underline{x}} f(\underline{x})$$

$$\text{subject to} \quad g_i(\underline{x}) \geq 0, \qquad i = 1,\ldots,m$$

The dual problem:

$$\min_{\underline{\lambda}} D(\underline{\lambda}) = \sup_{\underline{x}} \mathrm{L}(\underline{x}, \underline{\lambda}, \underline{\mu})$$

$$= \sup_{\underline{x}} f(\underline{x}) + \sum_{i=1}^{m} \lambda_i g_i(\underline{x})$$

An iterative method for the optimal solution to the dual problem:

$$\underline{x}^*(\underline{\lambda}(t)) = \arg\max L(\underline{x}, \underline{\lambda}(t)) = \arg\max \{f(\underline{x}) + \sum_{i=1}^{m} \lambda_i(t) g_i(\underline{x})\}$$

$$\lambda_i(t+1) = \lambda_i(t) - \delta_\lambda(t) g_i(\underline{x}^*(\underline{\lambda}(t)))$$

A sufficient condition for zero duality gap and that the iterations also yield the optimal solution for the primal problem:

If the price-based function $\underline{x}^*(\underline{\lambda}^*)$ is continuous around at least one of the optimal Lagrange multiplier vectors $\underline{\lambda}^*$

See Tychogiorgos, Gkelias, Leung (2013)

Inadequacy of Conventional Optimization for Resource Allocation

- Despite much effort, gradient-based iterative solutions may take time to converge
- Conventional approaches require precise system parameters
  - Parameter changes require independent re-run of optimization process
  - Optimization process may not provide robust performance for a given range of system parameters

Desirable to have a new, efficient and robust approach to solving constrained optimization problems

Imperial College London

Optimization by Machine Learning

Use Coupled LSTM Networks to Solve Constrained Optimization Problems

Zheyu (Joe) Chen*, Kin K. Leung*, Shiqiang Wang‡, Leandros Tassiulas§, Kevin Chan ¶, Don Towsley#

* Imperial College, London, UK
‡ IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
§ Yale University, New Haven, CT, USA
¶ U.S. Army Research Lab, Adelphi, MD, USA
# University of Massachusetts, Amherst, MA, USA

Constrained Optimization Problem and its Dual Problem

- Constrained optimization problem

$$\text{(P1)} \min_{x} f(x)$$

$$s.t.\ h(x) \leq 0$$

- By introducing the Lagrange multipliers λ, we form
  - Lagrange function

$$J(x, \lambda) = f(x) + \lambda\, h(x)$$

  - Dual optimization problem

$$\text{(P2)} \quad \max_{\lambda} \ \mathrm{J}(\arg\min_{x} J(x, \lambda), \lambda)$$

$$s.t.\ \lambda \geq 0$$

- According to the duality theory, P1 and P2 have the same optimal solution when the duality gap is zero

Imperial College London

Projection Function for Lagrange Multipliers to Avoid Numerical Issues

Assumption:
The strong duality holds (i.e., the duality gap is zero) for P1 and P2, and thus there exists at least a dual optimal λ^* and a primal optimal x^*

To satisfy $\lambda \geq 0$ and avoid numerical issues:
We define a "smooth" projection function $\psi(\lambda) \geq 0 \ \forall \lambda$ to form P3 as follows

(P1) $\min_x f(x)$

$s.t. \ h(x) \leq 0$

(P2) $\max_\lambda J(\arg\min_x J(x,\lambda), \lambda)$

$s.t. \ \lambda \geq 0$

(P3) $\max_\lambda J(x^*, \psi(\lambda))$

$s.t. \ x^* = \arg\min_x J\big(x, \psi(\lambda)\big)$

Theorem: Having λ^* as the optimal solution for P2 is equivalent to having $u^* = \psi(\lambda^*)$ as the optimal solution for P3.

The proposed CLSTMs aims to solve the optimal λ^* and x^* from P3

Solve the Optimization Problem by Coupled LSTMs

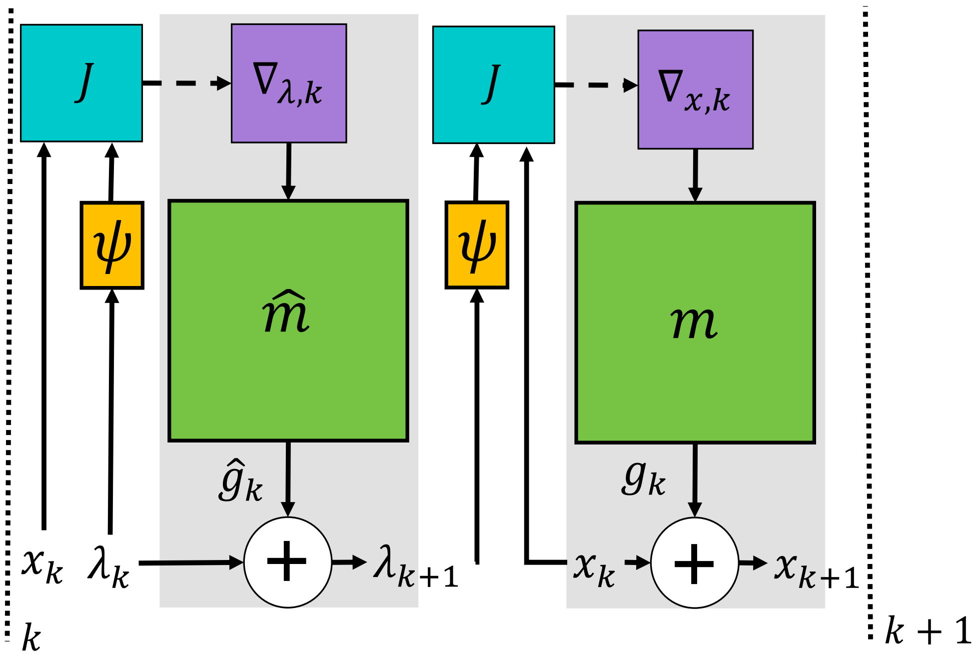
- During the inference process, the two coupled LSTMs, m and $\hat{m}$, are used to find the optimal x^* and λ^* by the following iterations:

$$\begin{bmatrix} \hat{g}_k \\ \hat{h}_{k+1} \end{bmatrix} = \hat{m}(\nabla_\lambda J(x_k, \psi(\lambda_k)), \hat{h}_k, \hat{\phi}),$$

$$\lambda_{k+1} = \lambda_k + \hat{g}_k,$$

$$\begin{bmatrix} g_k \\ h_{k+1} \end{bmatrix} = m(\nabla_x J(x_k, \psi(\lambda_{k+1})), h_k, \phi),$$

$$x_{k+1} = x_k + g_k,$$



$$\nabla_{\lambda,k} = \nabla_\lambda J(x_k, \psi(\lambda)) \quad \nabla_{x,k} = \nabla_x J(x_k, \psi(\lambda_{k+1}))$$

Training of the Coupled LSTMs

- In each iteration, x and λ are updated
- After K iterations (i.e., one frame), the parameters ϕ_i and $\hat{\phi}_i$ of the LSTMs m and $\hat{m}$ are updated to minimize the following loss functions:

$$L(\phi_i) = E_f[\sum_{k=(i-1)K}^{iK-1} w_k J(x_k, \psi(\lambda_{k+1})) + w_{iK} J(x_{iK}, \psi(\lambda_{iK})]$$

$$\hat{L}(\hat{\phi}_i) = -E_f[\sum_{k=(i-1)K}^{iK} \hat{w}_k J(x_k, \psi(\lambda_k))]$$

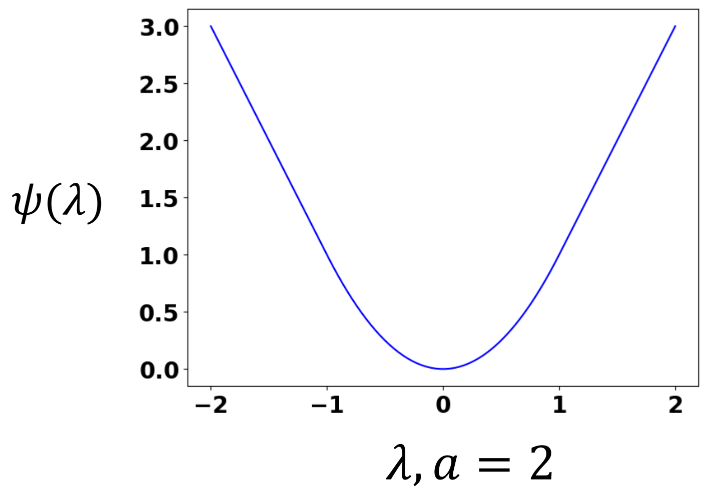
- At the start of every I frames, the variables $(x \text{ and } \lambda)$ and hidden states $(h_k \text{ and } \hat{h}_k)$ are randomly initialized

Imperial College London

Selection of Projection Function $\psi(\lambda)$ for Lagrange Multipliers

- To avoid numerical issues (e.g., calculating gradients), selection criteria for the projection function $\psi(\lambda)$ are:
  - $\psi(\lambda) \in [0, \infty)$ for all $\lambda \in \mathrm{R}$
  - $\psi(\lambda)$ is differentiable everywhere
  - When $\lambda \to \infty$ and $-\infty$, the derivatives of $\psi(\lambda)$ become non-zero constants, which can be different from 1 (i.e., $\psi(\lambda)$ increases or decreases linearly when λ is large)
  - The two constants should not be too small or large to avoid numerical issues
- An example of $\psi(\lambda)$:

$$\psi(\lambda) = \begin{cases} -a\lambda - (a-1), & \text{if } \lambda < -1 \\ \lambda^a, & \text{if } -1 \le \lambda \le 1 \\ a\lambda - (a-1), & \text{if } \lambda > 1 \end{cases}$$



Numerical Study: Resource Allocation

- The resource-allocation problem is to allocate cluster resources to competing jobs for maximizing the sum of job utilities

N	The number of jobs
C	The amount of available resource
r_n	The amount of resource allocated to the job n
R_n	The resource requirement of the job n
$u_n(r_n)$	The utility function given the allocated resource r_n
α, β	Two parameters to set the minimum and maximum amount of resource requirement of job n

$$\max_{r_1,\ldots,r_N} \sum_{n=1}^{N} u_n(r_n)$$

$$s.t. \sum_{n=1}^{N} r_n \leq C$$

$$r_n \geq \alpha R_n, \forall n$$

$$r_n \leq \beta R_n, \forall n$$

Experimental Setup

- Consider a cluster of 5 machines to provide CPU resource to 10 competing jobs
- In each problem scenario, the amount of available CPU resource and the CPU requirements of jobs are randomly selected from the Alibaba cluster trace

- Training process uses 5,120 problem scenarios
- Each LSTM of the CLSTMs has two layers and each layer has 20 neural units
- Proposed algorithm is implemented with Python and Tensorflow 2.1 and evaluated on an Ubuntu 20.04 LTS server with a NVIDIA TITAN XP graphics card

Imperial College London

Comparing the CLSTMs to Baseline Approaches

- Two Baseline Approaches for Comparison
  - Gradient descent (GD)
  - Gradient descent with momentum (GDM)
  - Baseline approach parameters are selected by exhaustively evaluating various parameter combinations
- Inference (evaluation) by the Trained CLSTMs
  - 1,000 problem scenarios
  - 2,000 iteration steps for each scenario
- Figure of Merit: Relative Accuracy to the True Optimum

$$\alpha = 1 - \frac{|\hat{f} - f|}{|f|}$$

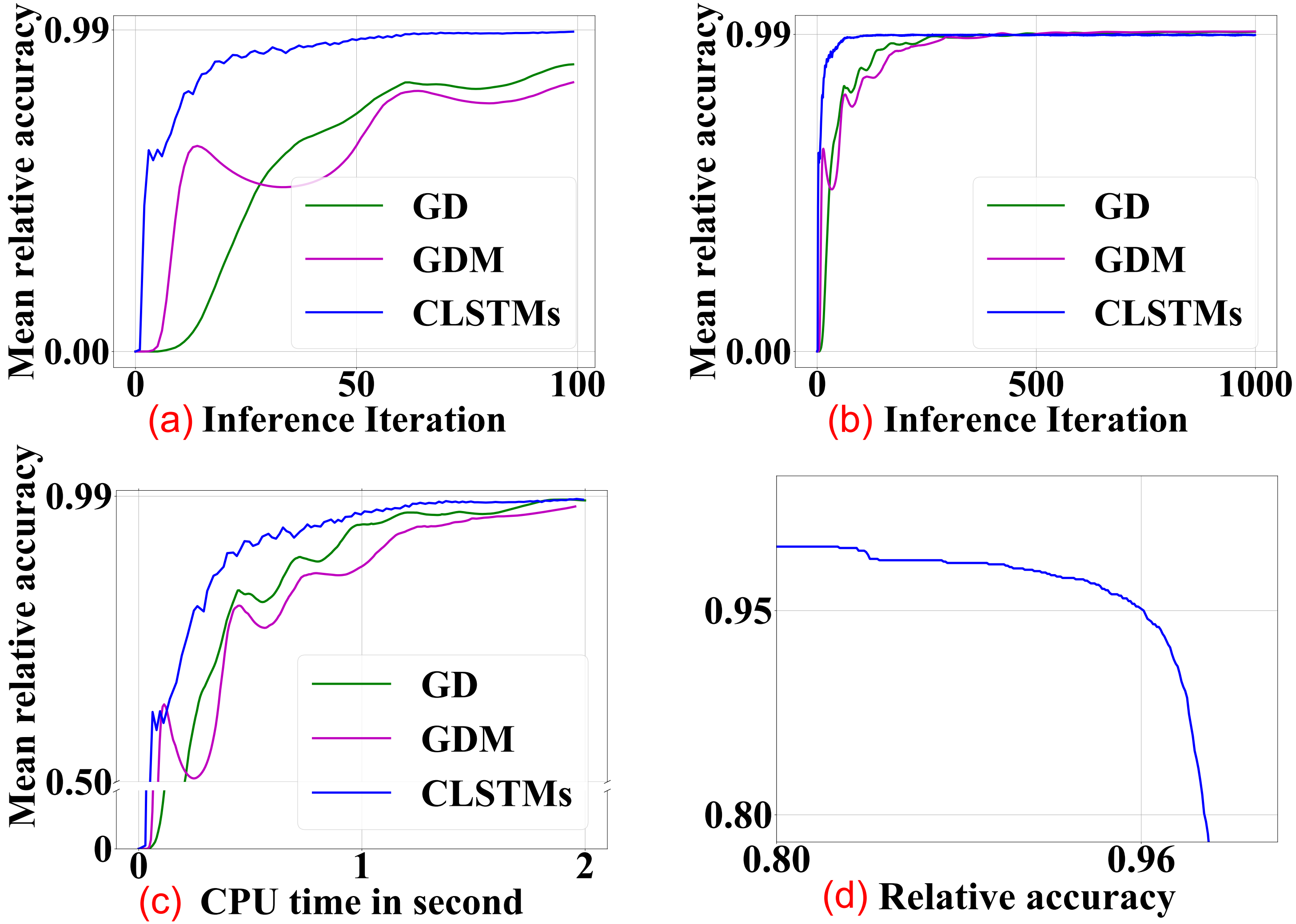
$\hat{f}$: the optimal objective function value found by the CLSTMs or baselines

f: the true optimal value of the objective function by the fmincon (i.e., in the Optimization-toolbox in Matlab R2016)

- Mean relative accuracy is the relative accuracy averaged over 1,000 problem scenarios

Significant Improvements by the CLSTMs

Mean relative accuracy over (a) 100 iterations, (b) 1,000 iterations, (c) CPU time in seconds, and (d) complementary cumulative distribution (CCDF) for relative accuracy



The number of iterations and CPU time consumed to achieve 90% mean relative accuracy are reduced by 86% and 56% relative to GDM, respectively

Imperial College
London

Impact of the projection functions

- Consider five projection functions $\psi(\lambda)$

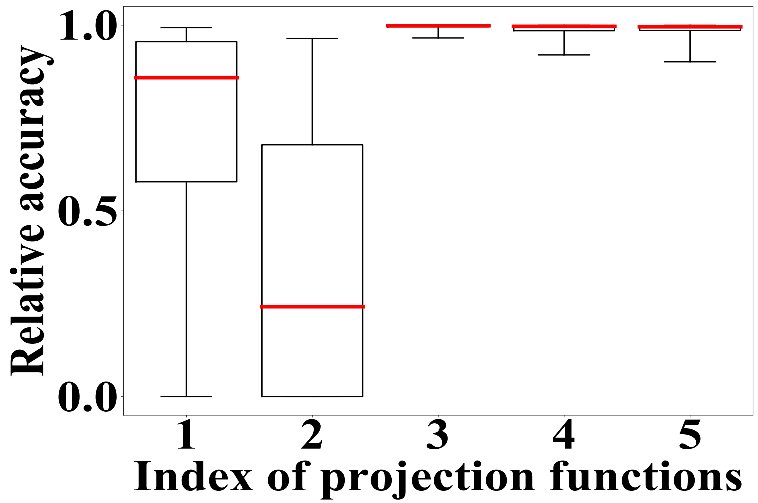
(1) $\psi(\lambda) = |\lambda|$

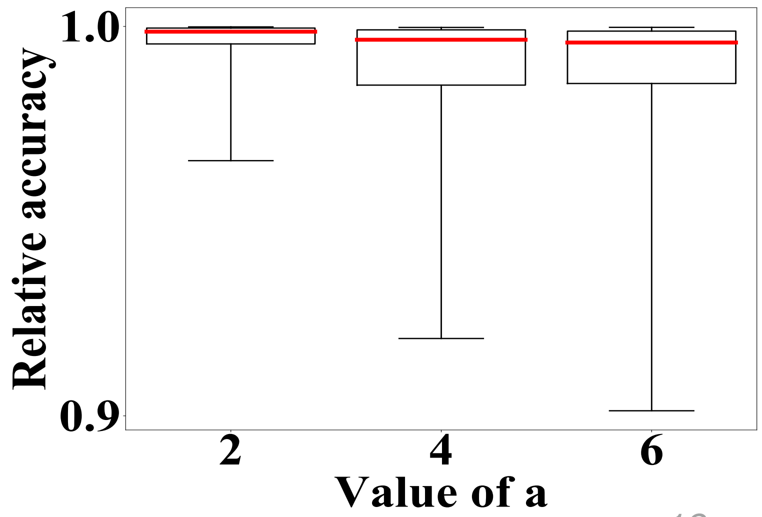
(2) $\psi(\lambda) = \frac{1}{2}\left(\sqrt{\lambda^2 + 0.25} + \lambda\right)$

$$\psi(\lambda) = \begin{cases} -a\lambda - (a-1), & if\ \lambda < -1 \\ \lambda^a, & if\ -1 \le \lambda \le 1 \\ a\lambda - (a-1), & if\ \lambda > 1 \end{cases}$$

(3) a=2, (4) a=4, and (5) a=6

- The lower whisker, the bottom of the box, the red horizontal line, the top of the box and the upper whisker represent the 5th, 25th, 50th, 75th and 95th percentile of the relative accuracy, respectively





Concluding Remarks and Future Direction

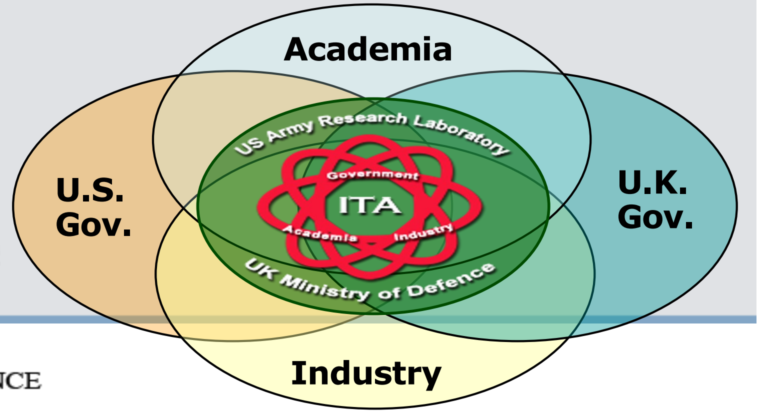
Concluding Remarks

- Optimization techniques have been shown to be helpful to resource and network management
- Nonconvex optimization and distributed solutions remain open
- Proposed a new machine-learning (ML) approach to solving constrained optimization problems
- ML approaches offer near-optimal and robust performance at a faster speed relative to conventional solution techniques

Future Research Direction

- Efficient management of network resources and services by solving optimization problems as quickly and accurately as by "distributed table lookup"!

Acknowledgments



- Acknowledgments
  - Zheyu Chen, Sepideh Nazemi, Faheem Zafari, George Tychogiorgos (Imperial College), Ananthram Swami and Kevin Chan (U.S. Army Research Lab), Shiqiang Wang (IBM), Leandros Tassiulas (Yale), Don Towsley (UMass) and Patrick Baker (UK Dstl/RAF)
  - Research funding: U.S./U.K. ITA Project

- Publications

Z. Chen, K.K. Leung, S. Wang, L. Tassiulas, K. Chan and D. Towsley, "Use Coupled LSTM Networks to Solve Constrained Optimization Problems," IEEE Trans. on Cognitive Communications and Networking, 2022.

Z. Chen, K.K. Leung, S. Wang, L. Tassiulas and K. Chan, "Robust Solutions to Constrained Optimization Problems by LSTM Networks," IEEE MILCOM 2021, pp. 503-508, Nov. 2021.

S. Nazemi, K.K. Leung and A. Swami, "Distributed Optimisation Framework for In-network Data Processing," *IEEE/ACM Transactions on Networking*, Vol. 27, No. 6, pp. 2432-2443, Dec. 2019.

F. Zafari, J. Li, K.K. Leung, D. Towsley and A. Swami, "Optimal Energy Consumption for Communication, Computation, Caching and Quality Guarantee," *IEEE Trans. on Control of Network Systems*, April 2019.

G. Tychogiorgos, A. Gkelias and K.K. Leung, "Distributed Network Resource Allocation for Multi-Tiered Multimedia Applications," IEEE INFOCOM, Hong Kong, China, April 2015.

G. Tychogiorgos and K.K. Leung, "Optimization-based Resource Allocation in Communication networks," *Computer Networks*, Vol. 66, pp. 32-45, June 2014.

G. Tychogiorgos, A. Gkelias and K.K. Leung, "A Non-Convex Distributed Optimization Framework and its Application to Wireless Ad-hoc Networks," *IEEE Trans. on Wireless Communications*, Vol. 12, pp. 4286 – 4296, September 2013.

Please google "Kin K Leung" for my website to find these and other papers.

Thank you