

Stochastic Optimization: Complexity-Based Analysis and Development Engineering Applications

Caleb Xavier Bugg
Operations Analyst

Reaching Our Sisters Everywhere, Inc.
Atlanta, GA (ROSE)

November 16, 2024

About me



About me



Outline

Logarithmic Bounds for Sample Average Approximation

- Sample Average Approximation (SAA)

- Rademacher Complexity

- Improved Sample Bounds

- Numerical Experiments

Nonnegative Tensor Completion via Integer Optimization

- The Tensor Completion Problem

- Past Approaches to TC

- Contributions: A New Norm for Nonnegative Tensors

- Results

Development Engineering: Optimal Intervention Theory (OIT)

- Global Poverty Alleviation and International Development (GPA & ID)

- Defining OIT

- Conclusions and Future Work

Outline I

Logarithmic Bounds for Sample Average Approximation

- Sample Average Approximation (SAA)

- Rademacher Complexity

- Improved Sample Bounds

- Numerical Experiments

Nonnegative Tensor Completion via Integer Optimization

- The Tensor Completion Problem

- Past Approaches to TC

- Contributions: A New Norm for Nonnegative Tensors

- Results

Development Engineering: Optimal Intervention Theory (OIT)

- Global Poverty Alleviation and International Development (GPA & ID)

- Defining OIT

- Conclusions and Future Work

Sample Average Approximation (SAA)

Sample Average Approximation (SAA) is a commonly-used procedure for approximating solutions to stochastic optimization problems of the form

$$\min_{x \in \mathcal{X}} \{F(x) := \mathbb{E}_{\xi} f(x, \xi)\}, \quad (1)$$

Sample Average Approximation (SAA)

Sample Average Approximation (SAA) is a commonly-used procedure for approximating solutions to stochastic optimization problems of the form

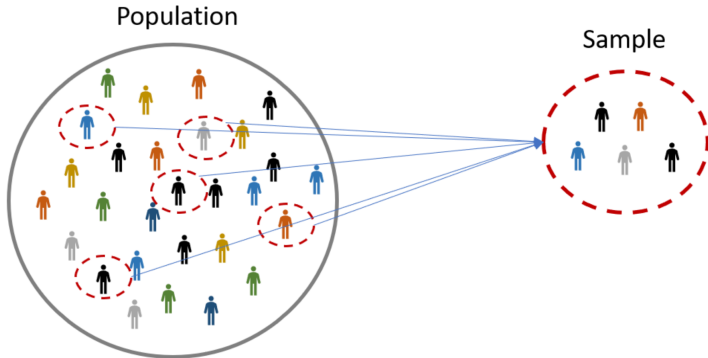
$$\min_{x \in \mathcal{X}} \{F(x) := \mathbb{E}_{\xi} f(x, \xi)\}, \quad (1)$$

The idea of SAA is to first generate an i.i.d. sample ξ_1, \dots, ξ_n of the random variable ξ , and then approximate the expectation $\mathbb{E}_{\xi} f(x, \xi)$ using its sample average

$$\min_{x \in \mathcal{X}} \{F_n(x) := \frac{1}{n} \sum_{i=1}^n f(x, \xi_i)\}. \quad (2)$$

Sample Bounds

The number of samples n in the SAA problem need to be as small as possible.



Sample Bounds

The number of samples n in the SAA problem need to be as small as possible.

Sample Bounds

The number of samples n in the SAA problem need to be as small as possible.

Towards this goal, a now classical analysis

[Kleywegt et al., 2002, Shapiro, 2003, Shapiro et al., 2009] showed that in order to ensure

$$\mathbb{P}(F(\hat{x}_n) - F(x^*) \leq \delta) \geq 1 - \alpha \quad (3)$$

for any $\delta \in (0, 1]$ and $\alpha \in (0, 1]$, the number of samples n should satisfy

$$n \gtrsim \frac{p}{\delta^2} \log \frac{1}{\delta} + \frac{1}{\delta^2} \log \frac{1}{\alpha}. \quad (4)$$

Sample Bounds

The number of samples n in the SAA problem need to be as small as possible.

Towards this goal, a now classical analysis

[Kleywegt et al., 2002, Shapiro, 2003, Shapiro et al., 2009] showed that in order to ensure

$$\mathbb{P}(F(\hat{x}_n) - F(x^*) \leq \delta) \geq 1 - \alpha \quad (3)$$

for any $\delta \in (0, 1]$ and $\alpha \in (0, 1]$, the number of samples n should satisfy

$$n \gtrsim \frac{p}{\delta^2} \log \frac{1}{\delta} + \frac{1}{\delta^2} \log \frac{1}{\alpha}. \quad (4)$$

These bounds depend polynomially on problem dimension p .



Rademacher Complexity

Analysis depends on stochastic process theory.

Rademacher Complexity

Analysis depends on stochastic process theory.

Let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher random variables, where $\mathbb{P}(\epsilon = \pm 1) = \frac{1}{2}$; and let $f(x, \xi)$ be the function from the objective of the SAA problem. We define the *Rademacher complexity* of the function set $\mathcal{F} := \{f(x, \xi) : x \in \mathcal{X}\}$ to be

$$\mathcal{R}_n[f] = \mathbb{E}_\xi \left(\sup_{x \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x, \xi_i) \right| \right). \quad (5)$$

Rademacher Complexity

With an assumption that $-\Delta/2 \leq f(x, \xi) \leq \Delta/2$ for all $(x, \xi) \in \mathcal{X} \times \Xi$, for some finite constant $\Delta \in \mathbb{R}_+$, we give a concentration bound of the form:

Proposition 1

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} |F_n(x) - F(x)| > t\right) \leq \exp\left(-2n\left(\frac{t - 2\mathcal{R}_n[f]}{\Delta}\right)^2\right). \quad (6)$$

The proof involves use of Jensen and McDiarmid's inequalities, a symmetrization argument, and an application of the triangle inequality.

Improved Sample Bounds

Proposition 2

Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be Lipschitz with constant L , and consider the stochastic optimization problem

$$\min_{x \in \mathcal{S}} \left\{ \mathbb{E}_{\xi} (g(\xi^T x)) \mid \|x\|_1 \leq \lambda \right\} \quad (7)$$

where $\mathcal{S} \subseteq \mathbb{R}^p$ and $\max_{\xi \in \Xi} \|\xi\|_{\infty} \leq C < +\infty$. Then the Rademacher complexity of the above problem is bounded by $\mathcal{R}_n[f] \leq \lambda LC \sqrt{2 \log 2p/n}$, and we need

$$n \geq \left(\frac{3\lambda LC}{\delta} \right)^2 \cdot \left(2 \log \left(\frac{2}{\alpha} \right) + 2 \log 2p \right) \quad (8)$$

samples to ensure that

$$\mathbb{P}(F(\hat{x}_n) - F(x^*) \leq \delta) \geq 1 - \alpha \quad (9)$$

for any $\delta \in (0, 1]$ and $\alpha \in (0, 1]$, holds.

Improved Sample Bounds

Proposition 3

Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be Lipschitz with constant L , and consider the stochastic optimization problem

$$\min_{X \in \mathcal{S}} \left\{ \mathbb{E}_{\xi} (g(\text{tr}(\xi^T X))) \mid \|X\|_* \leq \lambda \right\}$$

where $\mathcal{S} \subseteq \mathbb{R}^{p \times q}$ and $\max_{\xi \in \Xi} \|\xi\|_2 \leq C < +\infty$. Then the Rademacher complexity of the above stochastic optimization problem is bounded by

$\mathcal{R}_n[f] \leq \lambda LC \sqrt{3 \log(\min\{p, q\})/n}$, and we need

$$n \geq \left(\frac{3\lambda LC}{\delta} \right)^2 \cdot \left(2 \log \left(\frac{2}{\alpha} \right) + 3 \log (\min\{p, q\}) \right)$$

samples to ensure that for any $\delta \in (0, 1]$ and $\alpha \in (0, 1]$, (17) holds.

Numerical experiments

Consider a scenario where we would like to choose a portfolio that allocates investments into some combination of p risky assets and 1 risk-free asset, while considering a tradeoff between maximizing the expected return of the portfolio and the risk tolerance of the investor.



Numerical experiments

Consider a scenario where we would like to choose a portfolio that allocates investments into some combination of p risky assets and 1 risk-free asset, while considering a tradeoff between maximizing the expected return of the portfolio and the risk tolerance of the investor.



The Markowitz portfolio selection model [Markowitz, 1952, Bruder et al., 2013] is a simple framework to pose such a problem.

Numerical Experiments

Let $\xi \in \mathbb{R}^p$ is a random variable of the returns from the p risky assets, and define $\mu = \mathbb{E}_\xi \xi$ and $\Sigma = \mathbb{E}_\xi ((\xi - \mu)(\xi - \mu)^\top)$. Then one formulation of the problem involves solving a convex quadratic program

$$\min_{x \in \mathbb{R}^p} \left\{ x^\top \Sigma x - \gamma \cdot x^\top (\mu - r\mathbf{1}) \mid x \geq 0, \|x\|_1 \leq 1 \right\} \quad (10)$$

where:

r is the rate of return for the risk-free asset,

Numerical Experiments

Let $\xi \in \mathbb{R}^p$ is a random variable of the returns from the p risky assets, and define $\mu = \mathbb{E}_\xi \xi$ and $\Sigma = \mathbb{E}_\xi ((\xi - \mu)(\xi - \mu)^\top)$. Then one formulation of the problem involves solving a convex quadratic program

$$\min_{x \in \mathbb{R}^p} \left\{ x^\top \Sigma x - \gamma \cdot x^\top (\mu - r\mathbf{1}) \mid x \geq 0, \|x\|_1 \leq 1 \right\} \quad (10)$$

where:

r is the rate of return for the risk-free asset,

$\gamma > 0$ trades-off between the returns and risk of the portfolio, and

Numerical Experiments

Let $\xi \in \mathbb{R}^p$ is a random variable of the returns from the p risky assets, and define $\mu = \mathbb{E}_\xi \xi$ and $\Sigma = \mathbb{E}_\xi ((\xi - \mu)(\xi - \mu)^\top)$. Then one formulation of the problem involves solving a convex quadratic program

$$\min_{x \in \mathbb{R}^p} \left\{ x^\top \Sigma x - \gamma \cdot x^\top (\mu - r\mathbf{1}) \mid x \geq 0, \|x\|_1 \leq 1 \right\} \quad (10)$$

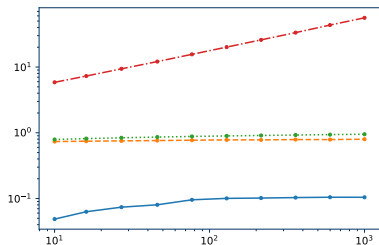
where:

r is the rate of return for the risk-free asset,

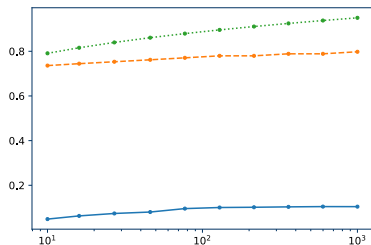
$\gamma > 0$ trades-off between the returns and risk of the portfolio, and

Each entry of the vector x gives the fraction of the portfolio allocated to the p risky assets; hence $1 - \sum_{i=1}^p x_i$ is the fraction of the portfolio allocated to the risk-free asset.

Numerical Experiments



(a) log-log plot



(b) semi-log plot

Figure 1: Comparison of 95% upper confidence bound of SAA solution gap (solid blue) with bounds on 95% upper confidence bound gap predicted classically (dash-dotted red), our Proposition (dashed orange), and our Corollary (dotted green).

In both plots, the x-axis is the dimension p of the decision variable, and the y-axis is the 95% upper confidence bound gap.

Outline I

Logarithmic Bounds for Sample Average Approximation

- Sample Average Approximation (SAA)

- Rademacher Complexity

- Improved Sample Bounds

- Numerical Experiments

Nonnegative Tensor Completion via Integer Optimization

- The Tensor Completion Problem

- Past Approaches to TC

- Contributions: A New Norm for Nonnegative Tensors

- Results

Development Engineering: Optimal Intervention Theory (OIT)

- Global Poverty Alleviation and International Development (GPA & ID)

- Defining OIT

- Conclusions and Future Work

Introduction

- Tensors generalize matrices.



Examples

Vectors are 1D tensors, matrices 2D, and so on...

Introduction

- Tensors generalize matrices.

Introduction

- Tensors generalize matrices.
- **Tensor Rank and Decomposition:** Though related, many problems that are polynomial-time solvable for matrices are NP-hard for tensors.

Introduction

- Tensors generalize matrices.
- **Tensor Rank and Decomposition:** Though related, many problems that are polynomial-time solvable for matrices are NP-hard for tensors.
- It is NP-hard to compute the rank of a tensor [Hillar and Lim, 2013], & tensor versions of the spectral norm, nuclear norm, and matrix singular value decomposition are also NP-hard to compute.
[Hillar and Lim, 2013, Friedland and Lim, 2014].

Tensor Completion, Past Approaches

Tensor completion is the problem of observing (possibly with noise) a subset of entries of a tensor and then estimating the remaining entries based on an assumption of low-rankness.

Tensor Completion, Past Approaches

Tensor completion is the problem of observing (possibly with noise) a subset of entries of a tensor and then estimating the remaining entries based on an assumption of low-rankness.

Suppose we have data $(x\langle i \rangle, y\langle i \rangle) \in \mathcal{R} \times \mathbb{R}$ for $i = 1, \dots, n$. Let $I = \{i_1, \dots, i_\nu\} \subseteq [n]$ be any set of points that specify all the unique $x\langle i \rangle$ for $i = 1, \dots, n$.

Tensor Completion, Past Approaches

Tensor completion is the problem of observing (possibly with noise) a subset of entries of a tensor and then estimating the remaining entries based on an assumption of low-rankness.

Suppose we have data $(x\langle i \rangle, y\langle i \rangle) \in \mathcal{R} \times \mathbb{R}$ for $i = 1, \dots, n$. Let $I = \{i_1, \dots, i_u\} \subseteq [n]$ be any set of points that specify all the unique $x\langle i \rangle$ for $i = 1, \dots, n$.

The nonnegative tensor completion problem is given by

$$\begin{aligned} \hat{\psi} \in \arg \min_{\psi} \frac{1}{n} \sum_{i=1}^n (y\langle i \rangle - \psi_{x\langle i \rangle})^2 \\ \text{s.t. } \text{rank}(\psi) \leq \lambda \end{aligned} \tag{11}$$

Tensor Completion, Past Approaches

Tensor completion is the problem of observing (possibly with noise) a subset of entries of a tensor and then estimating the remaining entries based on an assumption of low-rankness.

Suppose we have data $(x\langle i \rangle, y\langle i \rangle) \in \mathcal{R} \times \mathbb{R}$ for $i = 1, \dots, n$. Let $I = \{i_1, \dots, i_u\} \subseteq [n]$ be any set of points that specify all the unique $x\langle i \rangle$ for $i = 1, \dots, n$.

The nonnegative tensor completion problem is given by

$$\begin{aligned} \hat{\psi} \in \arg \min_{\psi} \frac{1}{n} \sum_{i=1}^n (y\langle i \rangle - \psi_{x\langle i \rangle})^2 \\ \text{s.t. } \text{rank}(\psi) \leq \lambda \end{aligned} \tag{11}$$

The “state of the art” computational methods use decomposition and an alternating minimization procedure to solve.

A “tension” in the TC World

Algorithms that achieve the information-theoretic rate have been developed for a few special cases of tensors.

A “tension” in the TC World

Algorithms that achieve the information-theoretic rate have been developed for a few special cases of tensors.

Completion of nonnegative rank-1 tensors can be written as a convex optimization problem [Aswani, 2016].

A “tension” in the TC World

Algorithms that achieve the information-theoretic rate have been developed for a few special cases of tensors.

Completion of nonnegative rank-1 tensors can be written as a convex optimization problem [Aswani, 2016].

For symmetric orthogonal tensors, a variant of the Frank-Wolfe algorithm has been proposed [Rao et al., 2015], which can be shown to achieve the information-theoretic rate.

A “tension” in the TC World

To date, no tensor completion algorithm has been shown to achieve the information-theoretic sample complexity rate, while guaranteeing convergence.

A “tension” in the TC World

To date, no tensor completion algorithm has been shown to achieve the information-theoretic sample complexity rate, while guaranteeing convergence.

Namely, for a tensor completion problem on a rank k tensor with sample size n , the information theoretic rate for estimation error is

$$\sqrt{k \cdot \sum_i r_i / n}$$

[Gandy et al., 2011].

A New Norm for Nonnegative Tensors

Our norm for nonnegative tensors uses a gauge (or Minkowski functional) construction, common in the Machine Learning world, and this provides some machinery for analysis [scaling of the ball].

A New Norm for Nonnegative Tensors

Our norm for nonnegative tensors uses a gauge (or Minkowski functional) construction, common in the Machine Learning world, and this provides some machinery for analysis [scaling of the ball].

It also depends on concepts and sets from the tensor world we cannot introduce here.

A New Norm for Nonnegative Tensors

Proposition 4

The function defined as

$$\|\psi\|_+ := \inf\{\lambda \geq 0 \mid \psi \in \lambda\mathcal{C}_1\} \quad (12)$$

is a norm for nonnegative tensors $\psi \in \mathbb{R}_+^{r_1 \times \dots \times r_p}$.

We will call the set \mathcal{C}_λ the nonnegative tensor polytope. A useful observation is that the following relationships hold: $\mathcal{B}_\lambda = \lambda\mathcal{B}_1$, $\mathcal{S}_\lambda = \lambda\mathcal{S}_1$, and $\mathcal{C}_\lambda = \lambda\mathcal{C}_1$.

A New Norm for Nonnegative Tensors

Proposition 4

The function defined as

$$\|\psi\|_+ := \inf\{\lambda \geq 0 \mid \psi \in \lambda\mathcal{C}_1\} \quad (12)$$

is a norm for nonnegative tensors $\psi \in \mathbb{R}_+^{r_1 \times \dots \times r_p}$.

We will call the set \mathcal{C}_λ the nonnegative tensor polytope. A useful observation is that the following relationships hold: $\mathcal{B}_\lambda = \lambda\mathcal{B}_1$, $\mathcal{S}_\lambda = \lambda\mathcal{S}_1$, and $\mathcal{C}_\lambda = \lambda\mathcal{C}_1$.

In our case \mathcal{C}_λ is not symmetric about the origin, and so without proof we do not *a priori* know whether scaling \mathcal{C}_1 eventually includes the entire space of nonnegative tensors. Thus we have to explicitly prove the gauge is a norm.

A New Norm for Nonnegative Tensors

Proposition 4

The function defined as

$$\|\psi\|_+ := \inf\{\lambda \geq 0 \mid \psi \in \lambda\mathcal{C}_1\} \quad (12)$$

is a norm for nonnegative tensors $\psi \in \mathbb{R}_+^{r_1 \times \dots \times r_p}$.

We will call the set \mathcal{C}_λ the nonnegative tensor polytope. A useful observation is that the following relationships hold: $\mathcal{B}_\lambda = \lambda\mathcal{B}_1$, $\mathcal{S}_\lambda = \lambda\mathcal{S}_1$, and $\mathcal{C}_\lambda = \lambda\mathcal{C}_1$.

In our case \mathcal{C}_λ is not symmetric about the origin, and so without proof we do not *a priori* know whether scaling \mathcal{C}_1 eventually includes the entire space of nonnegative tensors. Thus we have to explicitly prove the gauge is a norm.

Since the set of nonnegative tensors forms a cone [Qi et al., 2014], we must prove our norm using a modified definition of a norm (Proof omitted).

Results: Order 3 Tensors ($r = \text{dimensions}$)

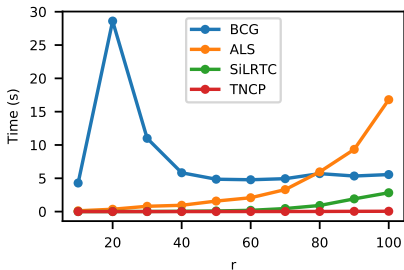
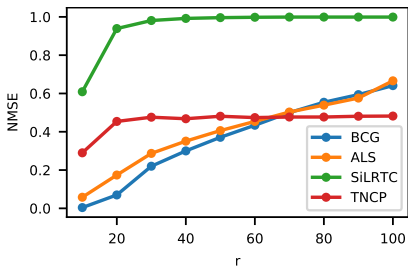


Figure 2: Results for order-3 nonnegative tensors with size $r \times r \times r$ and $n = 500$ samples.

Results: Increasing Tensor Order (p)

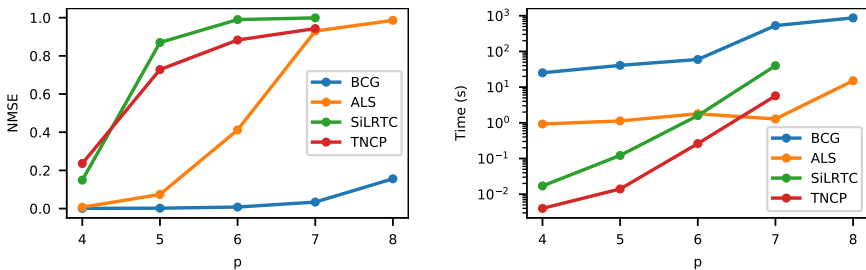


Figure 3: Results for increasing order nonnegative tensors with size $10^{\times p}$ and $n = 10,000$ samples.

Results: 10^6 entries and Increasing Sample Size

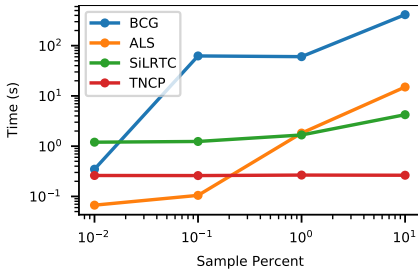
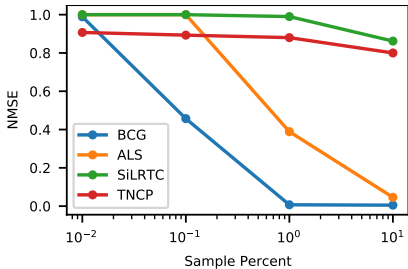


Figure 4: Results for nonnegative tensors with size 10^6 and increasing n samples.

Results: 10^7 entries and Increasing Sample Size

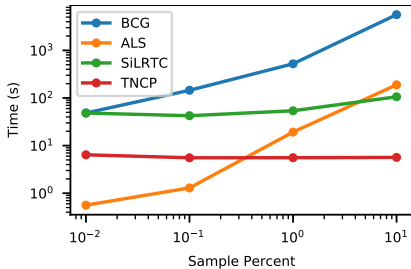
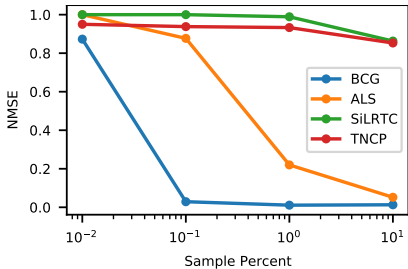


Figure 5: Results for nonnegative tensors with size $10^{\times 7}$ and increasing n samples.

Outline I

Logarithmic Bounds for Sample Average Approximation

- Sample Average Approximation (SAA)

- Rademacher Complexity

- Improved Sample Bounds

- Numerical Experiments

Nonnegative Tensor Completion via Integer Optimization

- The Tensor Completion Problem

- Past Approaches to TC

- Contributions: A New Norm for Nonnegative Tensors

- Results

Development Engineering: Optimal Intervention Theory (OIT)

- Global Poverty Alleviation and International Development (GPA & ID)

- Defining OIT

- Conclusions and Future Work



SUSTAINABLE DEVELOPMENT GOALS



UN SDGs



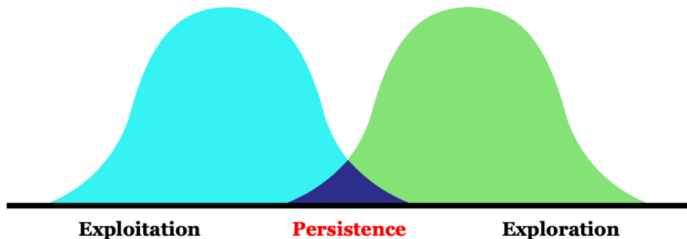
Optimal Intervention Theory (OIT)

Optimal Intervention Theory (OIT) is a method for improving human systems based on Statistical Learning Theory (SLT), which is the basis for learning in machines.

Optimal Intervention Theory (OIT)

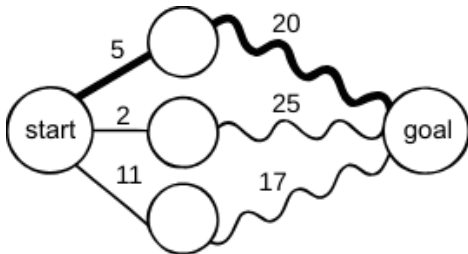
Optimal Intervention Theory (OIT) is a method for improving human systems based on Statistical Learning Theory (SLT), which is the basis for learning in machines.

Seeking a Productive Balance



A New Solution Framework: Optimal Intervention Theory (OIT)

- OIT considers the classic trade-off between statistically rigorous methods and large-scale methods
- The Dynamic Programming algorithm, which solves problems with well-defined end-goals in stages, is the basis of the SLT applicable to OIT.



Thank You

The full papers are available on my website at www.calebxb.com



References I



Aswani, A. (2016).

Low-rank approximation and completion of positive tensors.
SIAM Journal on Matrix Analysis and Applications, 37(3):1337–1364.



Bruder, B., Gaussel, N., Richard, J.-C., and Roncalli, T. (2013).

Regularization of portfolio allocation.
Available at SSRN 2767358.



Friedland, S. and Lim, L.-H. (2014).

Computational complexity of tensor nuclear norm.
Submitted.






Gandy, S., Recht, B., and Yamada, I. (2011).

Tensor completion and low-n-rank tensor recovery via convex optimization.
Inverse problems, 27(2):025010.

References II

-  Hillar, C. and Lim, L.-H. (2013).
Most tensor problems are np-hard.
J. ACM, 60(6):45:1–45:39.
-  Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. (2002).
The sample average approximation method for stochastic discrete optimization.
SIAM Journal on Optimization, 12(2):479–502.
-  Markowitz, H. (1952).
Portfolio selection.
Journal of Finance, 7(1):77–91.
-  Qi, Y., Comon, P., and Lim, L.-H. (2014).
Uniqueness of nonnegative tensor approximations.
arXiv preprint arXiv:1410.8129.

References III

-  Rao, N., Shah, P., and Wright, S. (2015).
Forward-backward greedy algorithms for atomic norm regularization.
IEEE Transactions on Signal Processing, 63(21):5798–5811.
-  Shapiro, A. (2003).
Monte Carlo sampling methods.
Handbooks in operations research and management Science, 10:353–425.
-  Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009).
Lectures on stochastic programming: modeling and theory.
SIAM.