Parameter Estimation in Large-Scale Optimization Models

John Birge
University of Chicago
Booth School of Business

International Colloquium on Stochastic Modeling and Optimization
Dedicated to the 80th birthday of Professor András Prékopa

General Theme

- As optimization models grow, so do the number of estimated or sampled parameters
- The chance of rare estimation events increases (close to 1)
- Optimization models are driven to extremes and naturally focus on “rare events” that slow convergence (or increase errors) and increase dependence on dimension
- *Challenge:* Can we find a way to avoid these problems? (Better ways to use available data?)

Example: Financial Portfolio Optimization

Quadratic program (Markowitz Portfolio):

find investments x=(x(1),…,x(n)) to

$$\min x^T Q x$$

$$\text{s.t. } r^T x = \text{target},\ e^T x=1$$

where Q and r are typically estimated from historical data.

Correlations from University of Michigan CIO:

	DomCommon	SmallCap	InteCommon	EmerMarkets	AbsoluteRetur	VentCap	RealEst	Oil and Gas	Commodities	FixedIncome	IntFixedInc
DomCommon	1	0.79	0.58	0.56	0.6	0.44	0.25	0.01	-0.3	0.43	0.2
SmallCap	0.79	1	0.48	0.61	0.65	0.56	0.24	0.01	-0.05	0.31	0.1
InteCommon	0.58	0.48	1	0.37	0.45	0.25	0.38	-0.04	-0.17	0.35	0.55
EmerMarkets	0.56	0.61	0.37	1	0.3	0.3	0.07	-0.19	-0.07	-0.07	0.1
AbsoluteRetur	0.6	0.65	0.45	0.3	1	0.35	0.2	-0.2	0.11	0.35	0.25
VentCap	0.44	0.56	0.25	0.3	0.35	1	0.21	-0.02	-0.18	0.19	0.15
RealEst	0.25	0.24	0.38	0.07	0.2	0.21	1	0.08	-0.53	0.15	0.2
Oil and Gas	0.01	0.01	-0.04	-0.19	-0.2	-0.02	0.08	1	0.54	-0.18	-0.3
Commodities	-0.3	-0.05	-0.17	-0.07	0.11	-0.18	-0.53	0.54	1	-0.3	-0.08
FixedIncome	0.43	0.31	0.35	-0.07	0.35	0.19	0.15	-0.18	-0.3	1	0.55
IntFixedInc	0.2	0.1	0.55	0.1	0.25	0.15	0.2	-0.3	-0.08	0.55	1
Cash	0.27	0.08	0.23	0.04	0.45	0.14	0.37	-0.07	-0.13	0.67	0.1

Results from Optimization

	Amt. to invest
DomCommon	-54079107483.07
SmallCap	-17314640179.88
InteCommon	-7098209713.34
EmerMarkets	21285151081.48
AbsoluteReturn	65911278495.65
VentCap	3346118938.17
RealEst	-68300117027.99
Oil and Gas	66227880616.79
Commodities	-104263997812.77
FixedIncome	-72656761795.57
IntFixedInc	117884874179.2
Cash	49057530702.32
Return	0.1
Variance	-1.65E+019

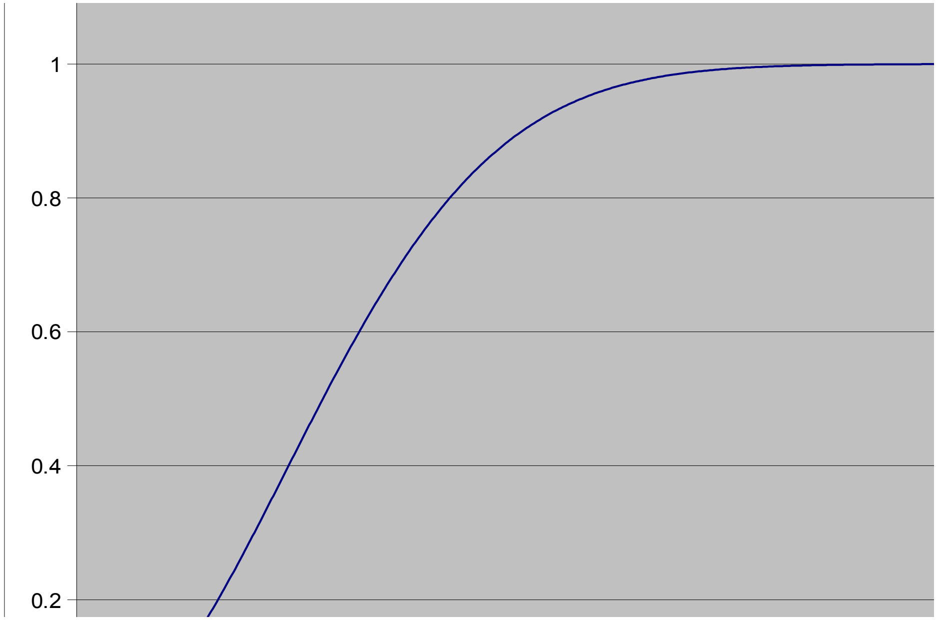
What happened here?

Problems in Markowitz Model

- Consistent time series
  - Correlations from different time series may not yield PD covariance matrices
  - Caution for general parameter estimates
- Number of Correlation Parameters
  - For n assets, $n(n-1)/2$ correlations to estimate
  - Chances of estimation error increase rapidly in n

Chance of Negative Correlation Observations

- Assume all true correlations are 3 standard deviations above 0 and each estimate is independent (not so but..)
- How does the probability of negative correlation observation relate to n (no. of assets)?



CHICAGO BOOTH

Problem Statement

- Large problems with n variables and m constraints/objective coefficients lead to (at least) mn estimates
- Probability of significant deviation from mean values increases rapidly in mn
- Deviant estimates drive optimal solutions
- ***How can we construct large models that yield consistent results with high probability?***

The General Questions

- Consider the basic problem (stochastic program):

$$\text{Min}_{x \varepsilon X} \; E_{\xi}[f(x,\xi)] \qquad\qquad (P)$$

- Suppose the only information for ξ is through samples: $\xi^1, \ldots, \xi^\nu$
- What can we say about solutions of sampled problems:

$$Min_{x\varepsilon\, X}\, (1/\,\nu) \sum_{i=1}^{\nu} f\,(x,\xi^i)$$

in relation to solution x^* to (P)?

- Are there better ways to use those samples?

General Sampling Result

(King-Rockafellar (1993, e.g.): Suppose $\boldsymbol{x}^{\nu}$ solves:

$\min_{x\in X} (1/\nu)\sum_{i=1}^{\nu} f(x,\xi^i)$

then, under a suitable set of conditions (X polyhedral, f smooth, unique optimum),

we can find a random vector, $\boldsymbol{u}$, that solves another optimization problem such that

$\nu^{0.5}(\boldsymbol{x}^{\nu} - x^*)$ converges to $\boldsymbol{u}$

Note: similar to a Central Limit Theorem but maybe even better. u is often Gaussian but often projected onto constraints.

Example of Asymptotic Distribution

- The asymptotic distribution of ***u*** depends on the constraints
- Example: Find x to

$$\min_{x \geq a} E[\| x - \xi \|]$$

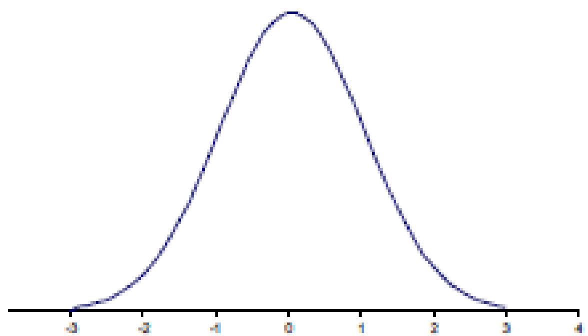
$$\text{where } \xi \sim N(0,1).$$

Note: $x^* = a$ for $a \geq 0$, 0 for a < 0.

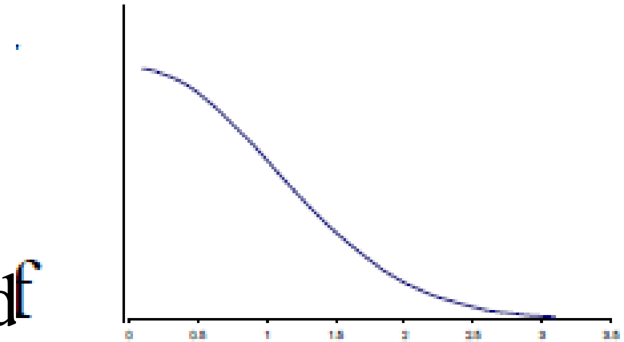
- What is the value of $u \sim \lim_{\nu} \nu^{0.5}(\mathbf{x}^{\nu} - x^*)$ for different a?

Possible Distributions for Example

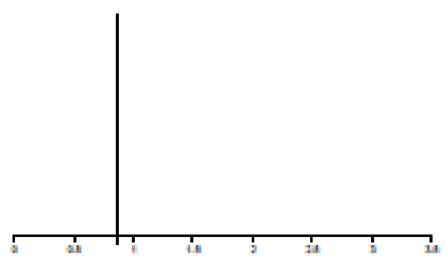
- $a < 0,\ u \sim N(0,1)$



- $a=0,\ P[u=0]=0.5$
  $F(u)=\Phi(u),\ u>0$
  for Φ normal cdf



- $a>0,\ u = 0$



Observations: The Good News

- Asymptotic distribution of optimal solution of sampled problem is:
  - Sometimes multivariate normal
  - Sometimes projection of multivariate normal onto constraints
  - Sometimes an atom at a single point
- Questions for large data sets:
  - When do we start to observe the asymptotic behavior?
  - How big must ν (no. of samples) be?

Quantitative Results

Goal: *Universal Confidence Sets* (e.g., Pflug (2003), Vogel (2008))

$$P\{|E_\xi[f(x^\nu,\xi)-f(x^*,\xi)]| \geq \epsilon\} \leq \alpha_1 e^{-\beta_1 \nu}.$$

and, if x^* *is unique,*

$$P\{\|x^\nu - x^*\| \geq \epsilon\} \leq \alpha_0 e^{-\beta_0 \nu}.$$

- Possible (sometimes explicit), e.g., Dai, Chen, JRB (2000)

Observations and Questions

- Have appealing asymptotic results that indicated confidence intervals might be possible
- Have universal bounds that indicate exponential convergence

Questions: 1. When do asymptotic properties appear? (Size of the constants?)

2. What are the effects of dimension? of multiple uncertainties? of constraints?

3. Are there better ways to use samples and, if so, when?

Form of Examples: Mean-Risk

Objective is composed of risk and return:

$E[f(x,w)] = - exp.return(x) + risk(x)$

For portfolios, often mean-variance, but can be different.

For uncertainty, sometimes only in the return, sometimes only in risk and sometimes in both – (this can effect convergence)

Example Problem

- Consider the following problem:

$min_x \; E_\xi[\, -\boldsymbol{\xi}^T x + \varepsilon \,||\, x \,||_1 \,]$

s. t. $-1 <= x <= 1$

where $||\,.||_1$ is the 1-norm (so equivalent to a linear program) and $E[\xi]=0$.

The optimal solution should be $x^*=0$.

How long to achieve limiting distribution?

How long will it take a sample solution to approach x^* *exponentially? i.e., when does* $Log\,(P\{||x^\nu - x^*|| \geq \varepsilon\}$ *decrease linearly?*

Sample Problem

- Assume that $\xi_j \sim N(0,1)$ for all j,

the solution is $x^{\nu}_{j} = 0$ if $|\xi_j| \leq \varepsilon$, and $\pm$ 1 o.w.

So, $P\{||x^{\nu}-x^{*}|| \geq 1\} = P\{|x^{\nu}_{j}| \geq 1$, some j$\}$

$= P\{|\xi_j| \geq \varepsilon$, some j$\} = 1-(1-2\Phi(-\gamma\nu^{0.5}))^{n}$

where Φ is the standard normal c.d.f.

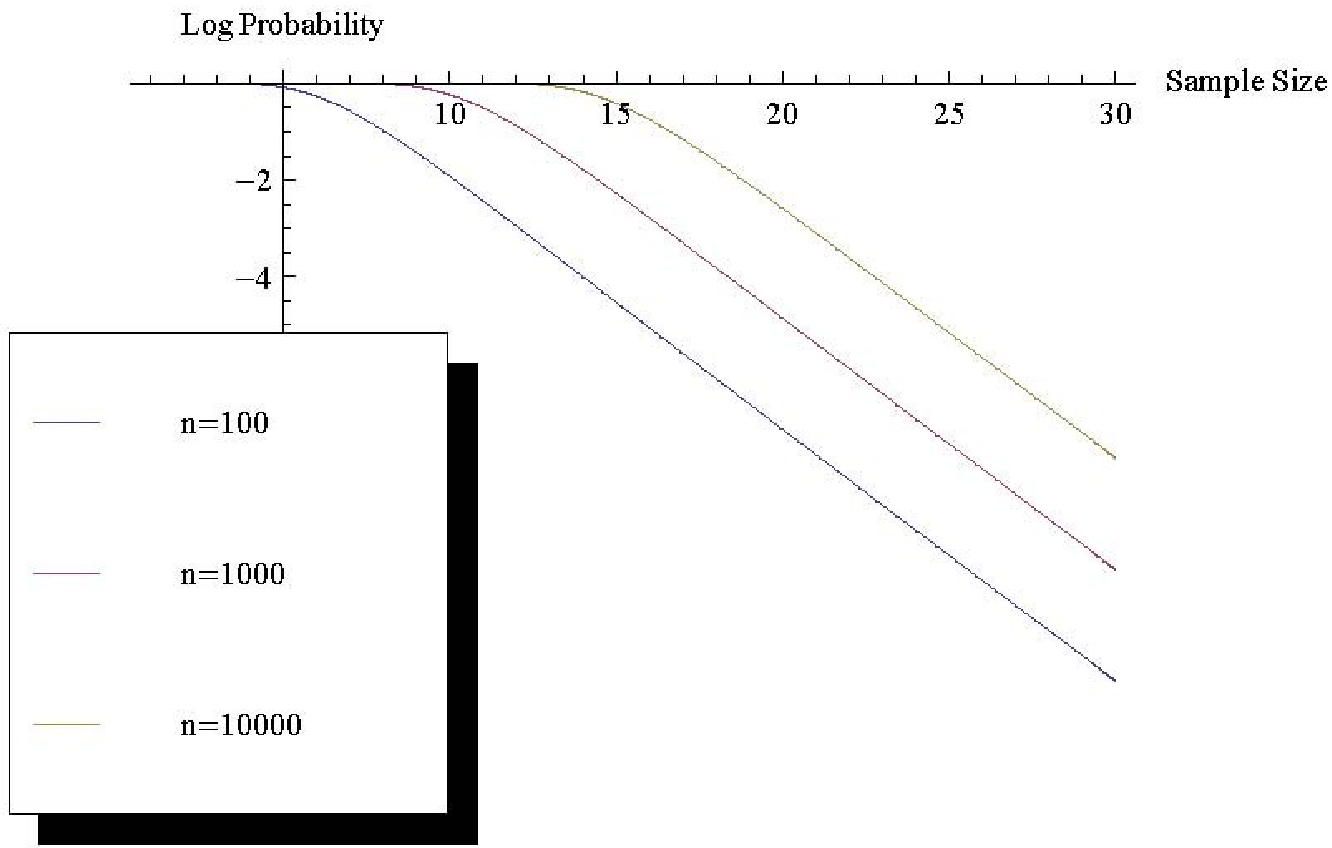
Note: already normal

When is Log(P(error≥1)) linear in ν?

What is the effect of dimension? (Note n)

Results

Log (P(error >1)) v. sample size (ν)



Observations

- Some delay in approach to exponential error decrease with dimension
- Increase in the delay (size of the constants in the universal bound) is less than linear in dimension (in fact, less than linear in Log of dimension)
- Same kinds of effects for objective
- Good results but could they be even better? Can we reduce the effect of the dimension?

How Can We Reduce the Required Number of Samples?

- Use of sub-samples or batch mean
- Suppose that we divide the ν samples into k batches of ν/k each, let ξ^{ν}_{i} be the mean of batch $i=1,\ldots,k$, then solve with ξ^{ν}_{i} to obtain x^{ν}_{i}
- Let $x^{\nu,k}=(1/k)\sum_{i=1}^{k} x^{\nu}_{i}$
- Can this do better?
- In particular, can we do better in the worst case?

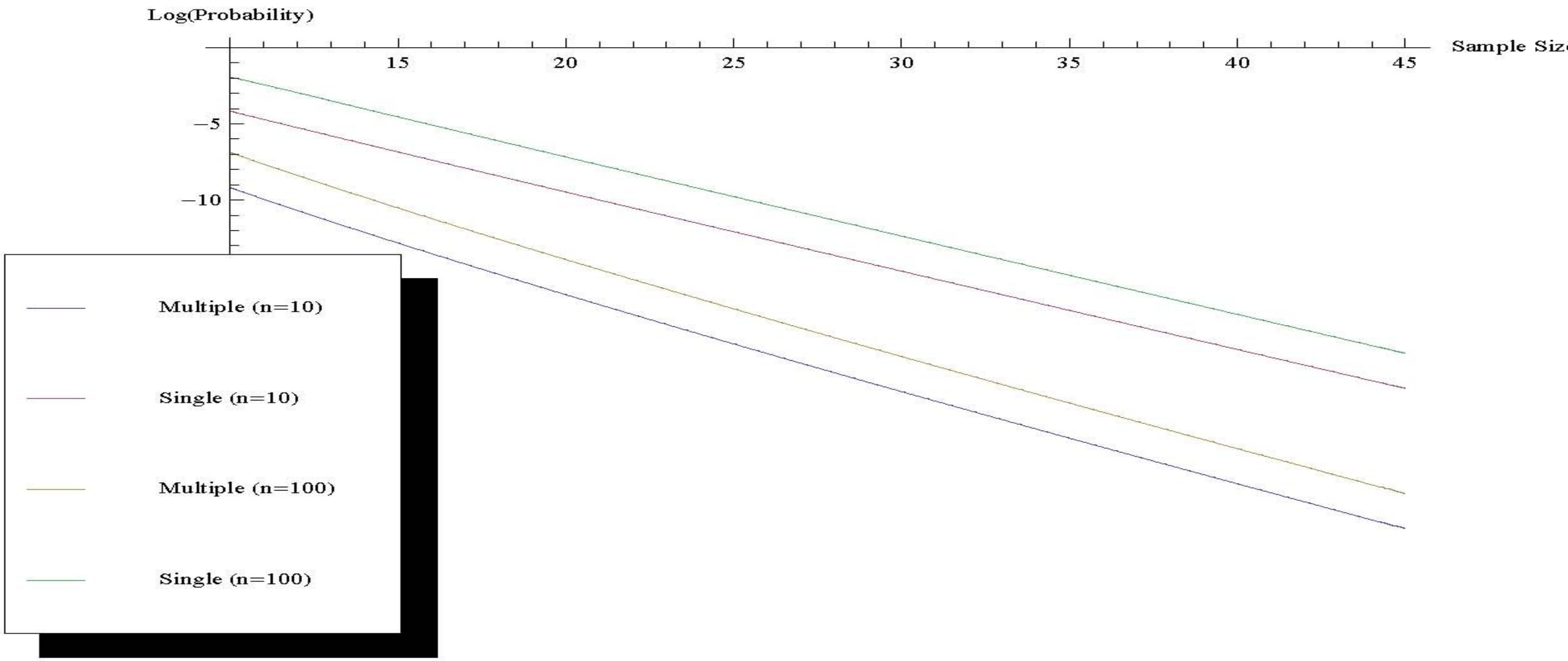
Result for Sub-sample Batch Optimization

- What is the chance that one component in the decision variable is far off?

$$P\{\|x^{\nu/K,K} - x^*\|_\infty \geq 1\} \leq P\{|x_j^{\nu,i}| \geq 1, \forall i = 1,\ldots,K; \text{for some } j \in \{1,\ldots,n\},\}$$
$$= 1 - (1 - (2\Phi(-\gamma(\nu/K)^{0.5}))^K)^n,$$

- Now, decreased dependence on *n*

Results for Batch/Single Samples



Observe: more improvement as ν ↑ (from 4 to 9 orders of magnitude)

What about Effects of Uncertainty in Risk?

- Example:

$$\min_{\|x\|_2 \le 1} E[-\xi^T x + \frac{\gamma}{2}\|x\|_2^2],$$

- Now, ξ and γ are random

Suppose $\xi_j \sim N(0,1)$; $\gamma \sim N(1,1)$

- Unconstrained solution:

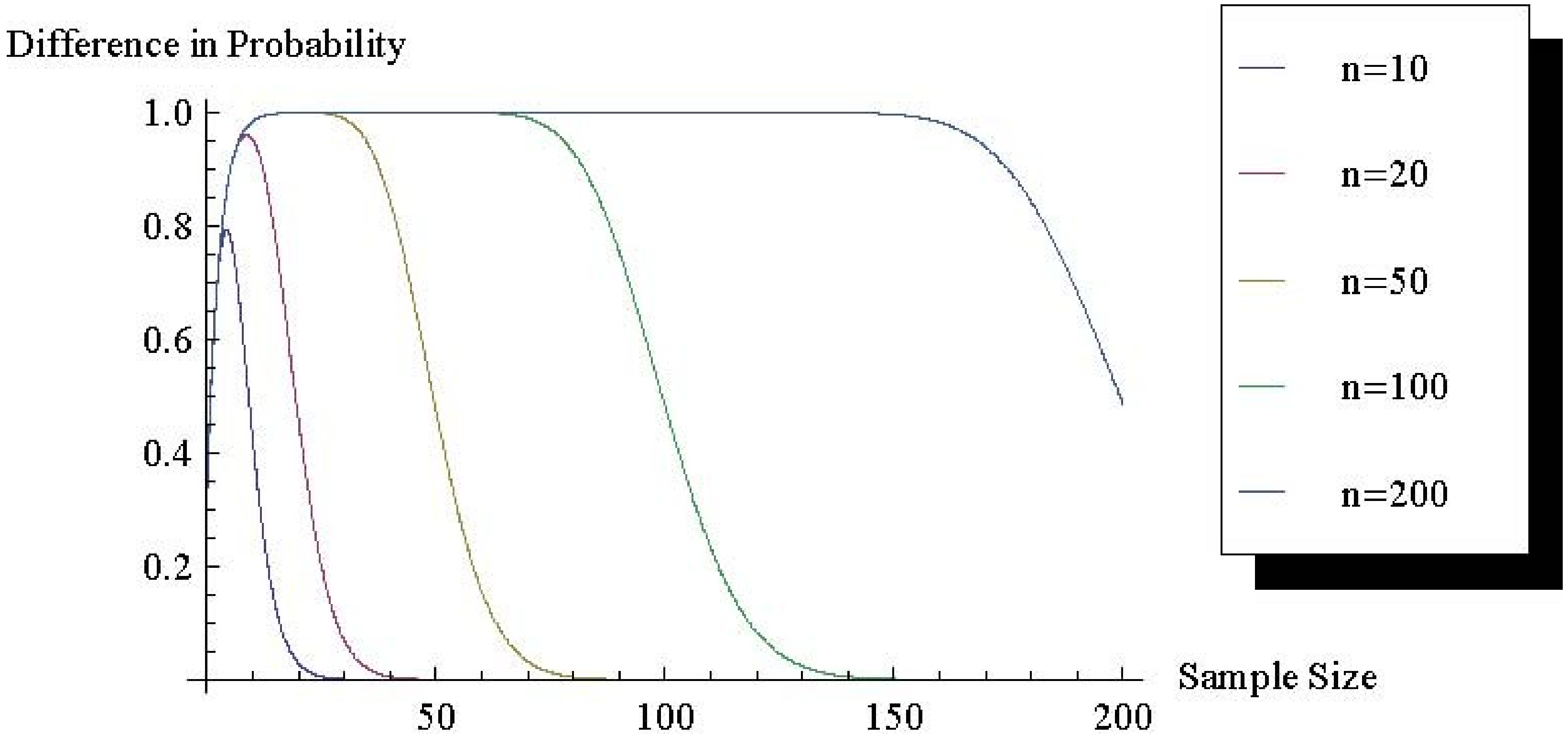
Error in solution in 2-norm is χ^2 under asymptotic distribution

True error in solution is given by:

$$\frac{1}{\|x^{\nu,u} - x^*\|_2^2} \sim F(1, n, \nu),$$

where F is the non-central F-ratio distribution

How Many Samples before the Error Approaches Asymptotic Distribution?



Observations

- Convergence now is much slower than in the case with just stochastic returns
- Convergence delay to the asymptotic distribution is almost linear in dimension
- Asymptotic distribution for the objective is again similar
- Asymptotic distribution for the general portfolio problem with multiple variance estimates (and inverse Wishart distribution) is even worse

Full Portfolio Examples

- General form:

$$\min_{x \in X} -\bar{r}^T x + \frac{\gamma}{2} x^T \Sigma x.$$

requires estimation: e.g., using sample estimates as:

$$\min_{x \in X} -\hat{r}^T x + \frac{\gamma(\nu - n - 2)}{2\nu} x^T \hat{\Sigma} x.$$

and $(\nu$-n-2$)/\nu$ term makes solution un-biased with no constraints (e.g., Kan and Zhou (2007))

Questions to Consider

- Can the use of sub-sample/batch optimal solutions improve convergence?
- How do the constraints affect the performance of the batch solution approximations?
- What is the effect of dimension in these problems?

Simulation Setup

For these results, we suppose $n = 10$, $\nu = 500$, and $K = 10$ and let $\gamma = 1$, $\mu = 0.2e$, where $e = (1, \ldots, 1)^T$, and $\Sigma = 0.05 * I$, where I is an identity matrix. We present the results from 1000 simulation runs for three different sets, X, corresponding to increasing ranges on x: $[0,1]^{10}$, $[-1,2]^{10}$, and $[-5,10]^{10}$. The results are compared relative to the optimal solution $x^* = 0.4e$ in terms of $\|x^\nu - x^*\|/\|x^*\|$ and optimal objective value $z^* = -\bar{r}^T x^* + \frac{1}{2} x^{*T} \Sigma x^* = -0.04$ in terms of $(-\bar{r}^T x^\nu + \frac{1}{2}(x^\nu)^T \Sigma x^\nu - z^*)/(-z^*)$.

Observe: histograms of relative errors in solutions and losses in objective

$X=[0,1]^{10}$

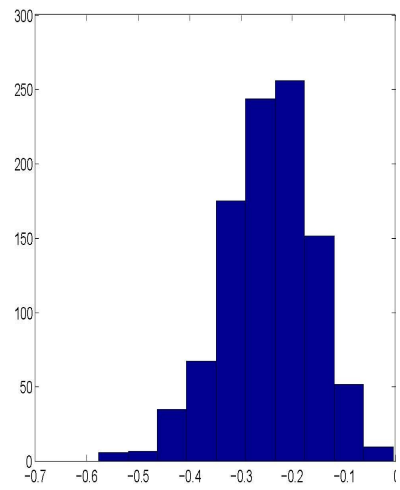
Relatives differences:

Batch better: 1000/1000

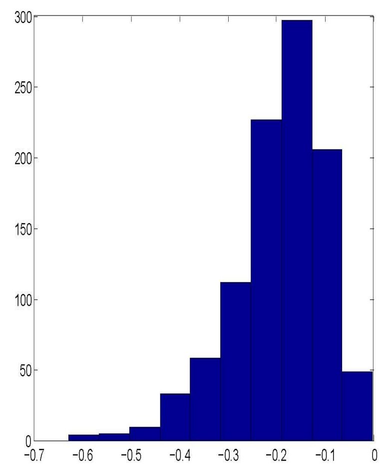
Avg. Sol. Dist. Diff. : -25%

Avg. Obj. Diff.: -19%

Solution



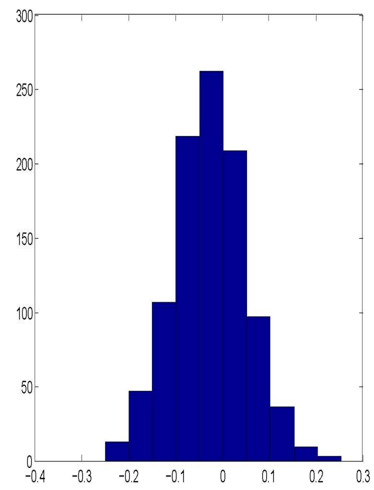
Objective

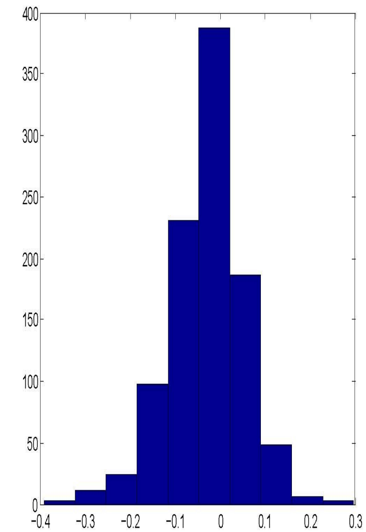


$X=[-1,2]^{10}$

Solution

Objective





Relatives differences:

Batch better: 638/1000

Avg. Sol. Dist. Diff. : -3%

Avg. Obj. Diff.: -3%

$X=[-5,10]^{10}$

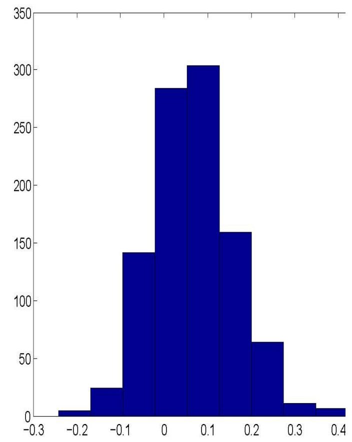
Relatives differences:

Batch better: 231/1000

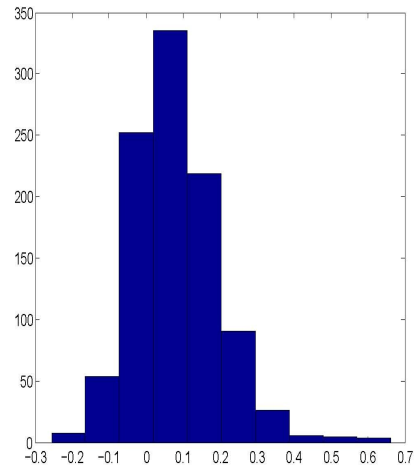
Avg. Sol. Dist. Diff. : +7%

Avg. Obj. Diff.: +8%

Solution



Objective

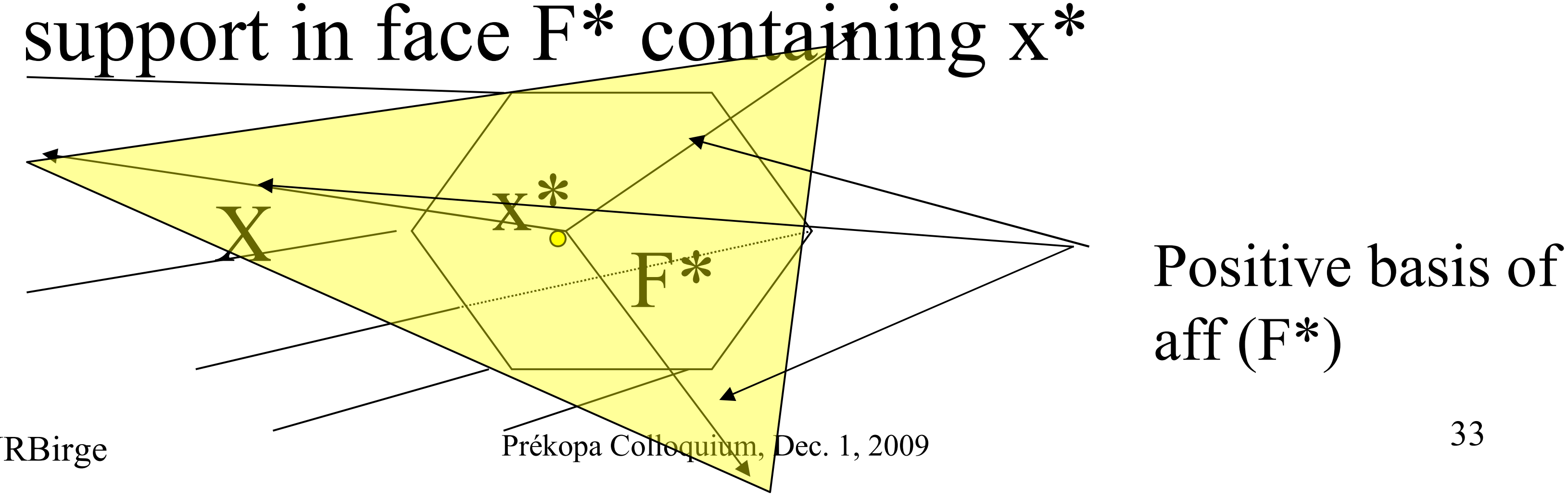


Observations on Portfolios

- Batch approach improves when constraints can bind the sample solutions
- The batch improvement is significant when constraints are relatively tight (but still more than 3 standard deviations from optimum)
- Batch can improve without constraints (but not so much in low dimensions ~10)

General Implications?

- How to put the batch results in terms of universal bounds?
- View: consider errors distributed throughout X and decompose by cone support in face F* containing x*



Assumptions

- Under mild conditions, x* is randomly distributed in F*
- Assume bias is known (or bounded)

$$b_{\nu/K} = \|E[x^{\nu/K} - x^*]\| \qquad O((\nu/K)^{-\frac{1}{2n}})$$

under certain regularity conditions (e.g., Roemisch and Schulz (1991))

- Worst error in any direction is g/n.

General Result

- Under these conditions,

$$P\left(\|\bar{u}^{\nu,K}\| \geq b_{\nu/K} + \frac{aM((N+1)g(N-g))^{1/2}}{K^{1/2}N}\right) \leq \frac{1}{a^2+1}.$$

- So, if unbiased, a=K$^{1/4}$,

error is greater than $\frac{g^{1/2}M}{K^{1/4}}$ with probability at most $\frac{1}{\sqrt{K}+1}$.

Implications of Result

- For relatively symmetric regions, the error from using batches can be of order even when asymptotics are not achieved within each sub-sample
- Non-symmetric regions may present difficulties (v→ 1/2n, worst case: isolated point)

F*:

Comparison to Other Approaches

- Imposing constraints (Jagannathan and Ma (2003))
- Shrinking variance (similar)
- "Re-sampled portfolio" (Michaud) - similar
- Robust optimization
- Bayesian updating
- Robust estimation
- Simple rules

Summary Observations

- Convergence to asymptotic behavior may be much slower with optimization and different uncertainty forms than simple estimation
- Dimension has more effect with greater uncertainty
- Use of optimization in batches can improve estimates especially with potentially violated constraints and symmetric feasible regions

Additional Questions

- Does the batch sample continue to improve with dimension in practical problems?
- Can these universal confidence sets be identified in the data?
- Are more general confidence interval estimates available?
- How do these approaches perform with other techniques to enhance convergence?

Thank You

and Happy Birthday, András!