

Optimization and Estimation in High Dimensions

John Birge
University of Chicago
Booth School of Business

General Themes

- 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:* What is the best way to use the available data?

Example: Financial Portfolio Optimization

Quadratic program (Markowitz Portfolio):

find investments $x=(x(1),\dots,x(n))$ to minimize risk (variance) subject to a constraint on expected return and using up a budget (normalized to 1):

$$\min x^T Q x$$

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

where Q and r are typically estimated from historical data.

Correlations from U. Michigan Chief Investment Officer:

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

Results from Optimization

	Amt. to invest
DomCommon	-54079107483
SmallCap	-17314640180
InteCommon	-7098209713
EmerMarkets	21285151081
AbsoluteReturn	65911278496
VentCap	3346118938
RealEst	-68300117028
Oil and Gas	66227880617
Commodities	-1.04264E+11
FixedIncome	-72656761796
IntFixedInc	1.17885E+11
Cash	49057530702
Return	0.099999487
Variance	-1.64591E+19

What happened here?



The General Questions

- Consider the basic problem (stochastic program):

$$\text{Min}_{x \in X} E_{\xi}[f(x, \xi)] \quad (P)$$

- Suppose the only information for ξ is through samples: ξ^1, \dots, ξ^v
- What is the best way to use those samples to find a solution x^* to (P)?

Sample Average Approximation

- Idea: use the samples as an empirical distribution with equal weight on each sample
- Solve to obtain x^o :

$$\min_{x \in X} \frac{1}{v} \sum_{i=1}^v f(x, \xi_i)$$

-*sample average approximation* (SAA)
problem (empirical loss in machine learning)

General Sampling Result

(King-Rockafellar (1993, e.g.): Suppose x^v solves:

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

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

we can find a random vector, u , that solves another optimization problem such that

$$v^{0.5}(x^v - x^*) \text{ converges to } u$$

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

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

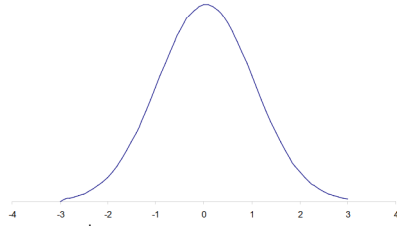
$$\text{where } \xi \sim N(0,1).$$

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

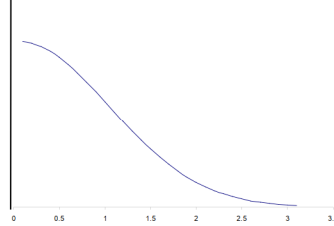
- What is the value of $u \sim \lim_v v^{0.5}(x^v - x^*)$ for different a ?

Possible Distributions for Example

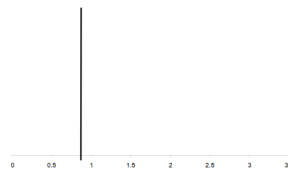
- $a < 0, u \sim N(0,1)$



- $a=0, P[u=0]=0.5$
 $F(u) = \Phi(u), u > 0$
 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:

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

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:

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

$$s. t. -1 \leq x \leq 1$$

where $\|\cdot\|_1$ 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 $\text{Log} (P\{\|x^v - x^\| \geq \varepsilon\})$ decrease linearly?*

Sample Problem

- Assume that $\xi_j \sim N(0,1)$ for all j , the solution is $x^v_j = 0$ if sample mean $\bar{\xi}_j^v$ has $|\bar{\xi}_j^v| \leq \epsilon$, and ± 1 o.w.

So, $P\{\|x^v - x^*\| \geq 1\} = P\{|x^v_j| \geq 1, \text{ some } j\}$
 $= P\{|\bar{\xi}_j^v| \geq \epsilon, \text{ some } j\} = 1 - (1 - 2\Phi(-\epsilon \sqrt{v}))^n$

where Φ is the standard normal c.d.f.

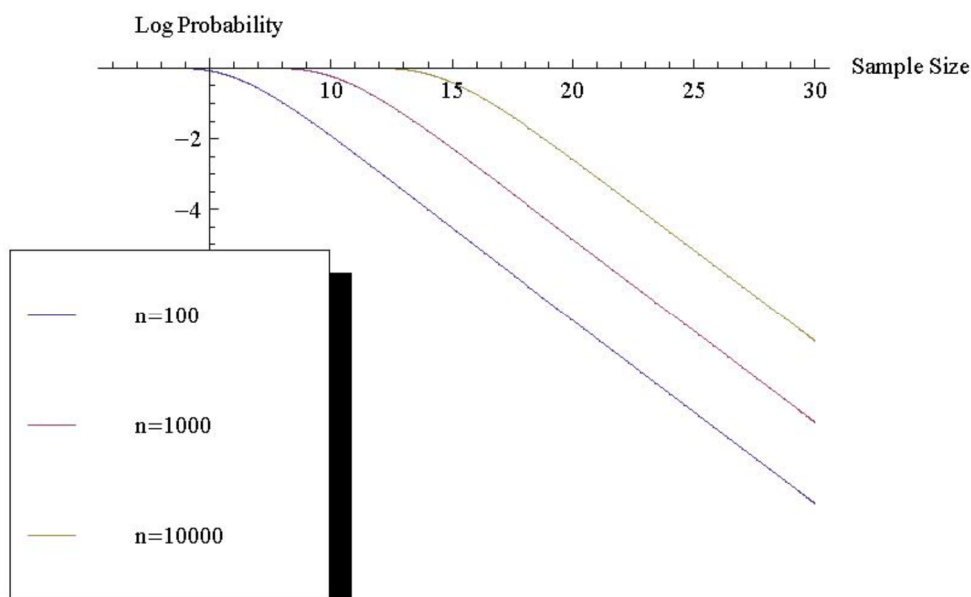
Note: already normal

When is $\text{Log}(P(\text{error} \geq 1))$ linear in v ?

What is the effect of dimension? (Note n)

Results

Log (P(error ≥ 1)) v. sample size (v)



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 (e.g., Mak, Morton, Wood (99))
- Suppose that we divide the ν samples into k batches of ν/k each, let $\xi^{\nu,l}$ be the mean of batch $i=1, \dots, k$, then solve with $\xi^{\nu,l}$ to obtain $x^{\nu,l}$
- Let $x^{\nu,k} = (1/k) \sum_{i=1}^k x^{\nu,l}$
- When does this perform better than a single sample?
- In particular, how much better in the worst case?

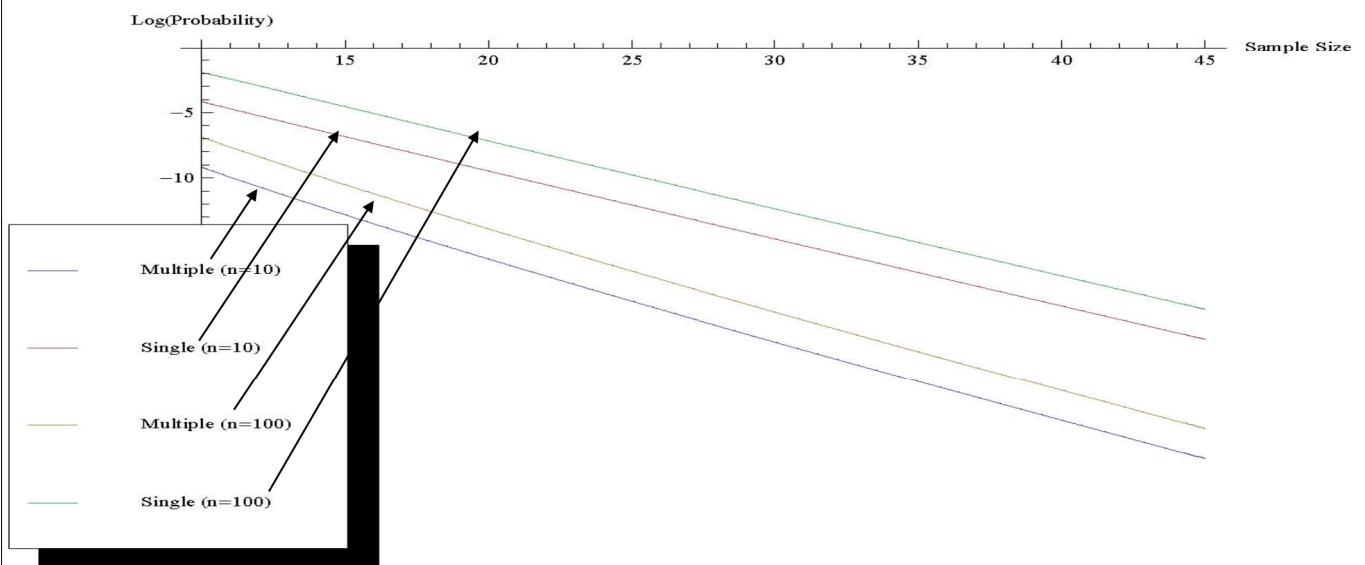
Result for Sub-sample Batch Optimization

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

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

- Now, decreased dependence on n

Results for Batch/Single Samples



Observe: more improvement as $\nu \uparrow$ (from 4 to 9 orders of magnitude)

What about Effects of Uncertainty in Risk?

- Example:

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

- 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

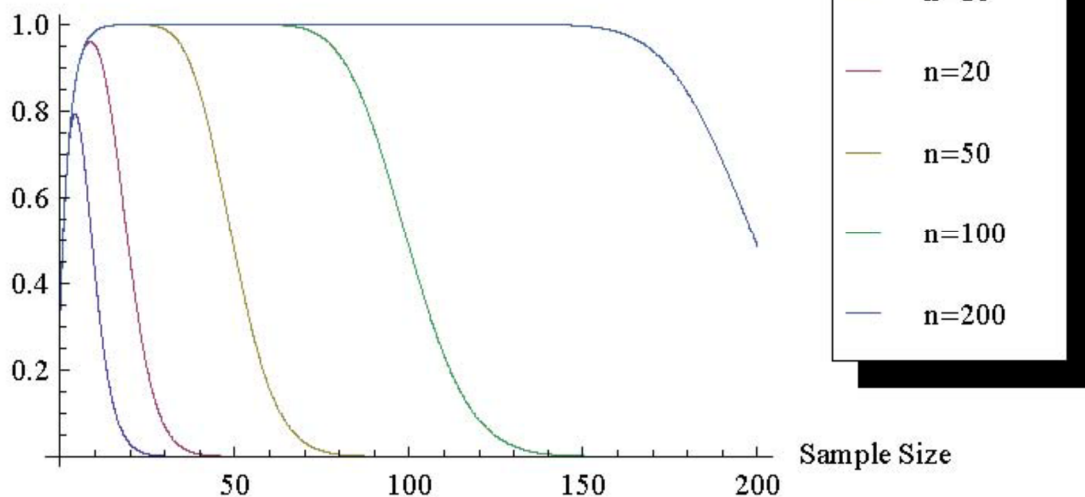
True error in solution is given by:

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

where F is the non-central F-ratio distribution

How Many Samples before the Error Approaches Asymptotic Distribution?

Difference in Probability



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

- Does 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 $e = (1, \dots, 1)^T$, 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

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

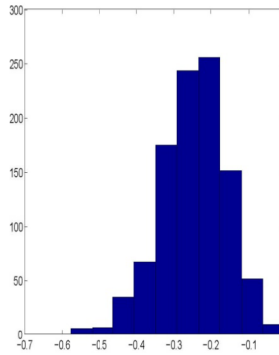
Relatives differences:

Batch better: 1000/1000

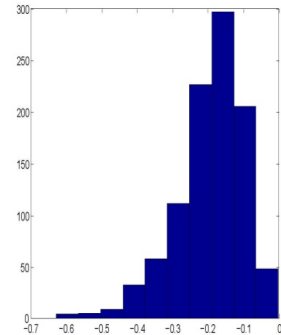
Avg. Sol. Dist. Diff. : -25%

Avg. Obj. Diff.: -19%

Solution



Objective



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

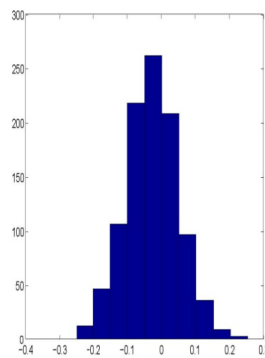
Relatives differences:

Batch better: 638/1000

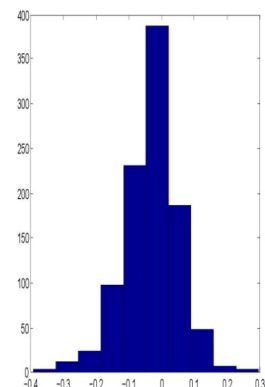
Avg. Sol. Dist. Diff. : -3%

Avg. Obj. Diff.: -3%

Solution



Objective



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

Relatives differences:

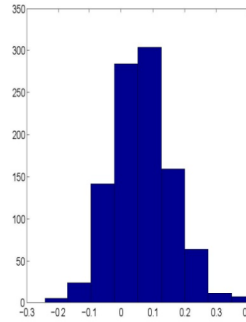
Batch better:
231/1000

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

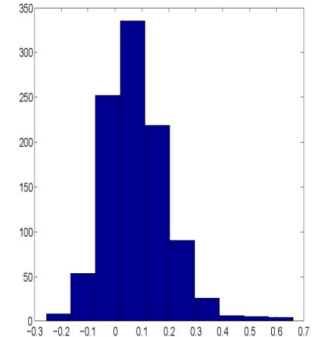
Avg. Obj. Diff.:
+8%

© JRBirge

Solution

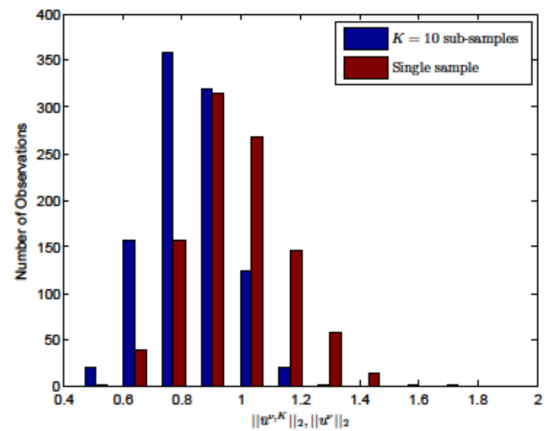
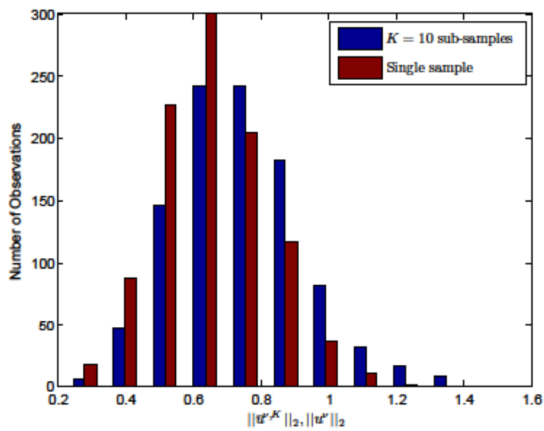


Objective



SoICT, HUST, Hanoi, 29 Aug 2022

10 to 20 dimensions



© JRBirge

SoICT, HUST, Hanoi, 29 Aug 2022

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)

Assumptions

- Under mild conditions, x^* is randomly distributed in optimal face F^* of dimension N
- Assume bias is known (or bounded), e.g.,

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

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

- Assume worst error in any direction of norm M is g/N

\Rightarrow Bound on error probability

General Result on Error Probability

- Under these conditions,

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

- So, if $b_{v/K} = 0, g \ll N, a = K^{\frac{1}{4}}$:

$$P\left(\text{error} \geq \frac{g^{1/2}M}{K^{1/4}}\right) \leq \frac{1}{\sqrt{K} + 1}$$

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? (Yes, in some way)
- Can these universal confidence sets be identified in the data? (Yes, if structure is known enough..)
- Are more general confidence interval estimates available?
- How do these approaches perform with other techniques to enhance convergence?

Thank you!