Decision Making Under Uncertainty: Lecture 2—Sample Average Approximation

Lecture 2

Ryan Cory-Wright

Spring 2026

Some Housekeeping

- Reminder: Please name the paper you are presenting for critical paper review and the week you are presenting in (by email to me) by Friday.

Some Housekeeping

- Reminder: Please name the paper you are presenting for critical paper review and the week you are presenting in (by email to me) by Friday.
- HW1 is now out, due on 2 Feb (see Insendi)—brief discussion of HW questions.
- I'll set aside some time at the end of the Monday Week 4 lecture, in case you have questions on the homework then.

Warm-Up: Solve This Problem

$$\begin{aligned} \min_{x_1, x_2} \quad & x_1 + x_2 \\ \text{s.t.} \quad & \omega_1 x_1 + x_2 \geq 7 \\ & \omega_2 x_1 + x_2 \geq 4 \\ & x_1, x_2 \geq 0 \end{aligned}$$

Where $\omega_1 \sim \mathcal{U}[1, 4], \omega_2 \sim \mathcal{U}[1/3, 1]$

Warm-Up: Solve This Problem

$$
\begin{aligned}
\min_{x_1, x_2} \quad & x_1 + x_2 \\
\text{s.t.} \quad & \omega_1 x_1 + x_2 \geq 7 \\
& \omega_2 x_1 + x_2 \geq 4 \\
& x_1, x_2 \geq 0
\end{aligned}
$$

Where $\omega_1 \sim \mathcal{U}[1, 4], \omega_2 \sim \mathcal{U}[1/3, 1]$

This problem is not well-enough defined to solve

Warm-Up: Solve This Problem

$$\begin{aligned}
\min_{x_1, x_2} \quad & x_1 + x_2 \\
\text{s.t.} \quad & \omega_1 x_1 + x_2 \geq 7 \\
& \omega_2 x_1 + x_2 \geq 4 \\
& x_1, x_2 \geq 0
\end{aligned}$$

Where $\omega_1 \sim \mathcal{U}[1, 4], \omega_2 \sim \mathcal{U}[1/3, 1]$
This problem is not well-enough defined to solve

First, we don't know how ω_1, ω_2 depend on each other.

Warm-Up: Solve This Problem

$$\begin{aligned}\min_{x_1,x_2} \quad & x_1 + x_2 \\ \text{s.t.} \quad & \omega_1 x_1 + x_2 \geq 7 \\ & \omega_2 x_1 + x_2 \geq 4 \\ & x_1, x_2 \geq 0\end{aligned}$$

Where $\omega_1 \sim \mathcal{U}[1,4], \omega_2 \sim \mathcal{U}[1/3,1]$ are independent

This problem is not well-enough defined to solve

First, we don't know how ω_1, ω_2 depend on each other. Assume independent

Warm-Up: Solve This Problem

$$\min_{x_1, x_2} \quad x_1 + x_2$$
$$\text{s.t.} \quad \omega_1 x_1 + x_2 \geq 7,\ \omega_2 x_1 + x_2 \geq 4,\ x_1, x_2 \geq 0$$

Where $\omega_1 \sim \mathcal{U}[1,4], \omega_2 \sim \mathcal{U}[1/3,1]$ are independent

This problem is not well-enough defined to solve

First, we don't know how ω_1, ω_2 depend on each other. Assume independent

Second, we don't know how x_1, x_2 depend on ω:

Warm-Up: Solve This Problem

$$\min_{x_1, x_2} \quad x_1 + x_2$$
$$\text{s.t.} \quad \omega_1 x_1 + x_2 \geq 7,\ \omega_2 x_1 + x_2 \geq 4,\ x_1, x_2 \geq 0$$

Where $\omega_1 \sim \mathcal{U}[1,4], \omega_2 \sim \mathcal{U}[1/3,1]$ are independent
This problem is not well-enough defined to solve

First, we don't know how ω_1, ω_2 depend on each other. Assume independent
Second, we don't know how x_1, x_2 depend on ω:

- Do we pick x, then Nature picks ω, or vice versa?

Warm-Up: Solve This Problem

$$\min_{x_1, x_2} \quad x_1 + x_2$$
$$\text{s.t.} \quad \omega_1 x_1 + x_2 \geq 7,\ \omega_2 x_1 + x_2 \geq 4,\ x_1, x_2 \geq 0$$

Where $\omega_1 \sim \mathcal{U}[1, 4], \omega_2 \sim \mathcal{U}[1/3, 1]$ are independent

This problem is not well-enough defined to solve

First, we don't know how ω_1, ω_2 depend on each other. Assume independent

Second, we don't know how x_1, x_2 depend on ω:

- Do we pick x, then Nature picks ω, or vice versa?
- First case: want to be feasible w.p.1., so minimizing $x_1 + x_2$ with $x_1 + x_2 \geq 7$, giving optimal solution of $(0, 7)$ with cost 7

Warm-Up: Solve This Problem

$$\min_{x_1, x_2} \quad x_1 + x_2$$

$$\text{s.t.} \quad \omega_1 x_1 + x_2 \geq 7,\ \omega_2 x_1 + x_2 \geq 4,\ x_1, x_2 \geq 0$$

Where $\omega_1 \sim \mathcal{U}[1,4], \omega_2 \sim \mathcal{U}[1/3,1]$ are independent

This problem is not well-enough defined to solve

First, we don't know how ω_1, ω_2 depend on each other. Assume independent

Second, we don't know how x_1, x_2 depend on ω:

- Do we pick x, then Nature picks ω, or vice versa?
- First case: want to be feasible w.p.1., so minimizing $x_1 + x_2$ with $x_1 + x_2 \geq 7$, giving optimal solution of $(0, 7)$ with cost 7
- Second case: more complicated casewise analysis (exercise)

Warm-Up: Solve This Problem

$$\min_{x_1, x_2} \quad x_1 + x_2$$
$$\text{s.t.} \quad \omega_1 x_1 + x_2 \geq 7,\ \omega_2 x_1 + x_2 \geq 4,\ x_1, x_2 \geq 0$$

Where $\omega_1 \sim \mathcal{U}[1,4], \omega_2 \sim \mathcal{U}[1/3, 1]$ are independent

This problem is not well-enough defined to solve

First, we don't know how ω_1, ω_2 depend on each other. Assume independent

Second, we don't know how x_1, x_2 depend on ω:

- Do we pick x, then Nature picks ω, or vice versa?
- First case: want to be feasible w.p.1., so minimizing $x_1 + x_2$ with $x_1 + x_2 \geq 7$, giving optimal solution of $(0, 7)$ with cost 7
- Second case: more complicated casewise analysis (exercise)

Conclusion: Terminology matters; should define everything carefully!

Outline of Lecture 2

Motivation: Ordinary Least Squares Regression

Linear Regression Setup—Rearranging

Linear regression: n i.i.d. observations of p-dimensional input vector $\boldsymbol{x}$ and output y, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$. We believe input-output follows model $y = \boldsymbol{x}^\top \boldsymbol{\beta}_{\text{true}} + \epsilon$, where $\boldsymbol{\beta}_{\text{true}}$ fixed vector, ϵ i.i.d. zero-mean noise.

Linear Regression Setup—Rearranging

Linear regression: n i.i.d. observations of p-dimensional input vector $\boldsymbol{x}$ and output y, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$. We believe input-output follows model $y = \boldsymbol{x}^\top \boldsymbol{\beta}_{\text{true}} + \epsilon$, where $\boldsymbol{\beta}_{\text{true}}$ fixed vector, ϵ i.i.d. zero-mean noise.

How to estimate $\boldsymbol{\beta}$?

Linear Regression Setup—Rearranging

Linear regression: n i.i.d. observations of p-dimensional input vector $\boldsymbol{x}$ and output y, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. We believe input-output follows model $y = \boldsymbol{x}^\top \boldsymbol{\beta}_{\text{true}} + \epsilon$, where $\boldsymbol{\beta}_{\text{true}}$ fixed vector, ϵ i.i.d. zero-mean noise.

How to estimate $\boldsymbol{\beta}$? Typical answer: minimize OLS error

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

Linear Regression Setup—Rearranging

Linear regression: n i.i.d. observations of p-dimensional input vector $\boldsymbol{x}$ and output y, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. We believe input-output follows model $y = \boldsymbol{x}^\top \boldsymbol{\beta}_{\text{true}} + \epsilon$, where $\boldsymbol{\beta}_{\text{true}}$ fixed vector, ϵ i.i.d. zero-mean noise.

How to estimate $\boldsymbol{\beta}$? Typical answer: minimize OLS error

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

After some calculus

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{y},$$

where $\boldsymbol{A}^\dagger$ denotes pseudoinverse of $\boldsymbol{A}$.

Linear Regression Setup—Rearranging

Linear regression: n i.i.d. observations of p-dimensional input vector $\boldsymbol{x}$ and output y, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. We believe input-output follows model $y = \boldsymbol{x}^\top \boldsymbol{\beta}_{\text{true}} + \epsilon$, where $\boldsymbol{\beta}_{\text{true}}$ fixed vector, ϵ i.i.d. zero-mean noise.

How to estimate $\boldsymbol{\beta}$? Typical answer: minimize OLS error

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

After some calculus

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{y},$$

where $\boldsymbol{A}^\dagger$ denotes pseudoinverse of $\boldsymbol{A}$. Assume p fixed, $n > p$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{y} \underbrace{=}_{\text{substitute } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{\text{true}} + \boldsymbol{\epsilon}} \boldsymbol{\beta}_{\text{true}} + (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{\epsilon}$$

Aside: Matrix Pseudoinverses

If $\boldsymbol{X}$ a matrix with singular value decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$
Then $\boldsymbol{X}^\dagger = \boldsymbol{V}\boldsymbol{\Sigma}^\dagger\boldsymbol{U}^\top$ where $\boldsymbol{\Sigma}^\dagger$ is a diagonal matrix where we invert all non-zero diagonal entries, keep zeroes as zeroes.

For a symmetric matrix like $\boldsymbol{X}^\top\boldsymbol{X}$, can define

$$(\boldsymbol{X}^\top\boldsymbol{X})^\dagger := \lim_{\lambda\to 0}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^\top.$$

See the book "Matrix Analysis" by Horn and Johnson.

Reminder: Almost Sure Convergence

Almost Sure Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\{\boldsymbol{X}_i\}_{i\in\mathbb{N}}, \boldsymbol{X}$ be random variables. Suppose that $\boldsymbol{A} \in \mathcal{F}$ is a measurable set such that $\mathbb{P}(\boldsymbol{A}) = 1$ and for all $\omega \in \mathcal{A}$ we have

$$\boldsymbol{X}_i(\omega) \to \boldsymbol{X}(\omega).$$

Then, we say that $\boldsymbol{X}_i \stackrel{a.s.}{\to} \boldsymbol{X}$.

Reminder: Continuous Mapping Theorem

Continuous Mapping Theorem

Let $\boldsymbol{X}_i, \boldsymbol{X}$ be random variables. Suppose that $\boldsymbol{X}_i \overset{a.s.}{\to} \boldsymbol{X}$ and f is continuous almost everywhere. Then

$$f(\boldsymbol{X}_i) \overset{a.s.}{\to} f(\boldsymbol{X})$$

Asymptotics of Linear Regression

Consider our rearranged equation:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{\beta}_{\text{true}} + (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{\epsilon}$$

As $n \to \infty$, what happens to $\hat{\boldsymbol{\beta}}$?

Asymptotics of Linear Regression

Consider our rearranged equation:

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{y} = \beta_{\text{true}} + (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \epsilon$$

As $n \to \infty$, what happens to $\hat{\beta}$?

- SLLN $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top \xrightarrow{a.s.} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top]$
- SLLN $\frac{1}{n}\boldsymbol{X}^\top \epsilon \xrightarrow{a.s.} \mathbf{0}$

Asymptotics of Linear Regression

Consider our rearranged equation:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{\beta}_{\text{true}} + (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{\epsilon}$$

As $n \to \infty$, what happens to $\hat{\boldsymbol{\beta}}$?

- SLLN $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top \overset{a.s.}{\to} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top]$
- SLLN $\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{\epsilon} \overset{a.s.}{\to} \mathbf{0}$
- Therefore $\hat{\boldsymbol{\beta}} \overset{a.s.}{\to} \boldsymbol{\beta}_{\text{true}}$ (under some mild conditions on span of $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top]$ etc.)

What did we just do?

- We solved our first stochastic optimization problem!

What did we just do?

- We solved our first stochastic optimization problem!
- Given sample of n data points $(\boldsymbol{x}_i, y_i)$, estimate model $\boldsymbol{\beta}$ by (1) writing down stochastic optimization problem

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{x},y}[(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2]$$

What did we just do?

- We solved our first stochastic optimization problem!
- Given sample of n data points $(\boldsymbol{x}_i, y_i)$, estimate model $\boldsymbol{\beta}$ by (1) writing down stochastic optimization problem

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{x},y}[(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2]$$

  find estimator with least variance, (2) treating each obs. as equally likely, replacing expectation with sample-average approximation

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

What did we just do?

- We solved our first stochastic optimization problem!
- Given sample of n data points $(\boldsymbol{x}_i, y_i)$, estimate model $\boldsymbol{\beta}$ by (1) writing down stochastic optimization problem

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{x},y}[(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2]$$

  find estimator with least variance, (2) treating each obs. as equally likely, replacing expectation with sample-average approximation

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

- We showed $\hat{\beta}$ almost surely converges to β_{true} as $n \to \infty$

What did we just do?

- We solved our first stochastic optimization problem!
- Given sample of n data points $(\boldsymbol{x}_i, y_i)$, estimate model $\boldsymbol{\beta}$ by (1) writing down stochastic optimization problem

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{x},y}[(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2]$$

  find estimator with least variance, (2) treating each obs. as equally likely, replacing expectation with sample-average approximation

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

- We showed $\hat{\beta}$ almost surely converges to β_{true} as $n \to \infty$
- So supervised learning is special case of stochastic optimization!

What did we just do?

- We solved our first stochastic optimization problem!
- Given sample of n data points $(\boldsymbol{x}_i, y_i)$, estimate model $\boldsymbol{\beta}$ by (1) writing down stochastic optimization problem

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{x},y}[(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2]$$

  find estimator with least variance, (2) treating each obs. as equally likely, replacing expectation with sample-average approximation

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

- We showed $\hat{\beta}$ almost surely converges to β_{true} as $n \to \infty$
- So supervised learning is special case of stochastic optimization!
- This would take a ML class 3-4 lectures; let's take a breath here!

What did we just do?

- We solved our first stochastic optimization problem!
- Given sample of n data points $(\boldsymbol{x}_i, y_i)$, estimate model $\boldsymbol{\beta}$ by (1) writing down stochastic optimization problem

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{x},y}[(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2]$$

find estimator with least variance, (2) treating each obs. as equally likely, replacing expectation with sample-average approximation

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

- We showed $\hat{\beta}$ almost surely converges to β_{true} as $n \to \infty$
- So supervised learning is special case of stochastic optimization!
- This would take a ML class 3-4 lectures; let's take a breath here!
- Plan for lecture: Show holds more generally, how to solve SAA

Sample Average Approximation: Theory

Let's warm up with a special case

Hot off the Press: The Newsvendor Problem

- A newsvendor (newspaper salesperson) needs to decide how many newspapers x to buy to maximize their profit

Hot off the Press: The Newsvendor Problem

- A newsvendor (newspaper salesperson) needs to decide how many newspapers x to buy to maximize their profit
- She doesn't know how many newspapers there are demand for, D_ω in scenario ω. But she does know the probability distribution of D_ω

Hot off the Press: The Newsvendor Problem

- A newsvendor (newspaper salesperson) needs to decide how many newspapers x to buy to maximize their profit
- She doesn't know how many newspapers there are demand for, D_ω in scenario ω. But she does know the probability distribution of D_ω
- Each newspaper costs c, can be sold for q if there is demand
- Unsold newspapers get thrown in the recycling bin
- How to optimally set x?

Hot off the Press: The Newsvendor Problem

$$\max_{x \geq 0} \mathbb{E}_\omega[\min(D_\omega, x)q - cx]$$

Hot off the Press: The Newsvendor Problem

$$\max_{x \geq 0} \mathbb{E}_\omega[\min(D_\omega, x)q - cx]$$

Two cases: $x > D_\omega$ or $x \leq D_\omega$. Rewrite using conditional expectations

Hot off the Press: The Newsvendor Problem

$$\max_{x \geq 0} \mathbb{E}_\omega[\min(D_\omega, x)q - cx]$$

Two cases: $x > D_\omega$ or $x \leq D_\omega$. Rewrite using conditional expectations

$$\max_{x \geq 0} \mathbb{E}_\omega[D_\omega q - cx | x \leq D_\omega]\mathbb{P}(x \leq D_\omega) + \mathbb{E}_\omega[qx - cx | x > D_\omega]\mathbb{P}(x > D_\omega)$$

Hot off the Press: The Newsvendor Problem

$$\max_{x \geq 0} \mathbb{E}_\omega[\min(D_\omega, x)q - cx]$$

Two cases: $x > D_\omega$ or $x \leq D_\omega$. Rewrite using conditional expectations

$$\max_{x \geq 0} \mathbb{E}_\omega[D_\omega q - cx | x \leq D_\omega]\mathbb{P}(x \leq D_\omega) + \mathbb{E}_\omega[qx - cx | x > D_\omega]\mathbb{P}(x > D_\omega)$$

This is convex in x, so differentiate with respect to x, require that 0 in subgradient.

Hot off the Press: The Newsvendor Problem

$$\max_{x \geq 0} \mathbb{E}_\omega[\min(D_\omega, x)q - cx]$$

Two cases: $x > D_\omega$ or $x \leq D_\omega$. Rewrite using conditional expectations

$$\max_{x \geq 0} \mathbb{E}_\omega[D_\omega q - cx | x \leq D_\omega]\mathbb{P}(x \leq D_\omega) + \mathbb{E}_\omega[qx - cx | x > D_\omega]\mathbb{P}(x > D_\omega)$$

This is convex in x, so differentiate with respect to x, require that 0 in subgradient.

Eventually get

$$x^\star \in F_{D_\omega}^{-1}\left(\frac{q - c}{q}\right)$$

Hot off the Press: The Newsvendor Problem

$$\max_{x \geq 0} \mathbb{E}_\omega[\min(D_\omega, x)q - cx]$$

Two cases: $x > D_\omega$ or $x \leq D_\omega$. Rewrite using conditional expectations

$$\max_{x \geq 0} \mathbb{E}_\omega[D_\omega q - cx | x \leq D_\omega]\mathbb{P}(x \leq D_\omega) + \mathbb{E}_\omega[qx - cx | x > D_\omega]\mathbb{P}(x > D_\omega)$$

This is convex in x, so differentiate with respect to x, require that 0 in subgradient.

Eventually get

$$x^\star \in F_{D_\omega}^{-1}\left(\frac{q-c}{q}\right)$$

That is, a $\frac{(q-c)}{q}$th quantile of D_ω

Insight: setting x equal to $\mathbb{E}[D_\omega]$ could be bad, especially if $q \gg c$

The General Problem

Overall Problem Setting: Two-Stage Stochastic Linear Opt

Consider stochastic optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \mathbb{E}_\omega[h(\boldsymbol{x}, \boldsymbol{\omega})] \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \end{aligned}$$

Overall Problem Setting: Two-Stage Stochastic Linear Opt

Consider stochastic optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \mathbb{E}_\omega[h(\boldsymbol{x}, \boldsymbol{\omega})] \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

where

$$\begin{aligned} h(\boldsymbol{x}, \boldsymbol{\omega}) := \quad & \min_{\boldsymbol{y}(\boldsymbol{\omega})} \boldsymbol{q}(\boldsymbol{\omega})^\top \boldsymbol{y}(\boldsymbol{\omega}) \\ \text{s.t.} \quad & \boldsymbol{D}(\omega)\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{\omega})\boldsymbol{y}(\boldsymbol{\omega}) \leq \boldsymbol{d}(\boldsymbol{\omega}) \end{aligned}$$

Overall Problem Setting: Two-Stage Stochastic Linear Opt

Consider stochastic optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \mathbb{E}_\omega[h(\boldsymbol{x}, \boldsymbol{\omega})] \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

where

$$\begin{aligned} h(\boldsymbol{x}, \boldsymbol{\omega}) := \quad & \min_{\boldsymbol{y}(\boldsymbol{\omega})} \boldsymbol{q}(\boldsymbol{\omega})^\top \boldsymbol{y}(\boldsymbol{\omega}) \\ \text{s.t.} \quad & \boldsymbol{D}(\omega)\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{\omega})\boldsymbol{y}(\boldsymbol{\omega}) \leq \boldsymbol{d}(\boldsymbol{\omega}) \end{aligned}$$

- x are our first-stage (or here-and-now) decision variables, which we select before nature picks $\boldsymbol{\omega}$

Overall Problem Setting: Two-Stage Stochastic Linear Opt

Consider stochastic optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \mathbb{E}_\omega[h(\boldsymbol{x},\boldsymbol{\omega})] \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \end{aligned}$$

where

$$\begin{aligned} h(\boldsymbol{x},\boldsymbol{\omega}) := \quad & \min_{\boldsymbol{y}(\boldsymbol{\omega})} \boldsymbol{q}(\boldsymbol{\omega})^\top \boldsymbol{y}(\boldsymbol{\omega}) \\ \text{s.t.} \quad & \boldsymbol{D}(\omega)\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{\omega})\boldsymbol{y}(\boldsymbol{\omega}) \le \boldsymbol{d}(\boldsymbol{\omega}) \end{aligned}$$

- x are our first-stage (or here-and-now) decision variables, which we select before nature picks $\boldsymbol{\omega}$
- $\boldsymbol{\omega}$ are the random variables selected by nature, according to their joint probability distribution (assumed to be known)

Overall Problem Setting: Two-Stage Stochastic Linear Opt

Consider stochastic optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \mathbb{E}_\omega[h(\boldsymbol{x}, \boldsymbol{\omega})] \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

where

$$\begin{aligned} h(\boldsymbol{x}, \boldsymbol{\omega}) := \quad & \min_{\boldsymbol{y}(\boldsymbol{\omega})} \boldsymbol{q}(\boldsymbol{\omega})^\top \boldsymbol{y}(\boldsymbol{\omega}) \\ \text{s.t.} \quad & \boldsymbol{D}(\omega)\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{\omega})\boldsymbol{y}(\boldsymbol{\omega}) \leq \boldsymbol{d}(\boldsymbol{\omega}) \end{aligned}$$

- x are our first-stage (or here-and-now) decision variables, which we select before nature picks $\boldsymbol{\omega}$
- $\boldsymbol{\omega}$ are the random variables selected by nature, according to their joint probability distribution (assumed to be known)
- $y(\boldsymbol{\omega})$ are our second-stage (or wait-and-see, or recourse) decision variables, that we are allowed to pick after nature picks $\boldsymbol{\omega}$

Overall Problem Setting: Two-Stage Stochastic Linear Opt

Consider stochastic optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \mathbb{E}_\omega[h(\boldsymbol{x}, \boldsymbol{\omega})] \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \end{aligned}$$

where

$$\begin{aligned} h(\boldsymbol{x}, \boldsymbol{\omega}) := \quad & \min_{\boldsymbol{y}(\boldsymbol{\omega})} \boldsymbol{q}(\boldsymbol{\omega})^\top \boldsymbol{y}(\boldsymbol{\omega}) \\ \text{s.t.} \quad & \boldsymbol{D}(\omega)\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{\omega})\boldsymbol{y}(\boldsymbol{\omega}) \le \boldsymbol{d}(\boldsymbol{\omega}) \end{aligned}$$

- x are our first-stage (or here-and-now) decision variables, which we select before nature picks $\boldsymbol{\omega}$
- $\boldsymbol{\omega}$ are the random variables selected by nature, according to their joint probability distribution (assumed to be known)
- $y(\boldsymbol{\omega})$ are our second-stage (or wait-and-see, or recourse) decision variables, that we are allowed to pick after nature picks $\boldsymbol{\omega}$
- A linear optimization problem with random parameters

What Makes This Problem Hard?

- Complexity Theory: Solving this problem is $\#P$-hard

What Makes This Problem Hard?

- Complexity Theory: Solving this problem is $\#P$-hard
  - See Hanasusanto et al. (Math. Prog., 2016) for a proof

What Makes This Problem Hard?

- Complexity Theory: Solving this problem is $\#P$-hard
  - See Hanasusanto et al. (Math. Prog., 2016) for a proof
  - Who knows what this means?

What Makes This Problem Hard?

- Complexity Theory: Solving this problem is $\#P$-hard
  - See Hanasusanto et al. (Math. Prog., 2016) for a proof
  - Who knows what this means?
  - As hard as counting number of solutions to NP-hard problem

What Makes This Problem Hard?

- Complexity Theory: Solving this problem is $\#P$-hard
    - See Hanasusanto et al. (Math. Prog., 2016) for a proof
    - Who knows what this means?
    - As hard as counting number of solutions to NP-hard problem
    - I once heard someone say "Judging a problem by its complexity is like judging someone by the worst thing they have ever done"
      —In and of itself, $\#P$-hard doesn't mean intractable

What Makes This Problem Hard?

- Complexity Theory: Solving this problem is $\#P$-hard
  - See Hanasusanto et al. (Math. Prog., 2016) for a proof
  - Who knows what this means?
  - As hard as counting number of solutions to NP-hard problem
  - I once heard someone say "Judging a problem by its complexity is like judging someone by the worst thing they have ever done" —In and of itself, $\#P$-hard doesn't mean intractable
- Structure of Optimal Solutions: In general, $\boldsymbol{y}$ a function of ω

Sample Average Approximation to the Rescue

Let's play same game as in the linear regression case!

Sample Average Approximation to the Rescue

Let's play same game as in the linear regression case!
Replace (unknown) expectation over $\boldsymbol{\omega}$ with expectation over empirical distribution $\boldsymbol{\omega}_i$. With n observations of $\boldsymbol{\omega}$, or n "scenarios", solve:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \end{aligned}$$

Sample Average Approximation to the Rescue

Let's play same game as in the linear regression case!

Replace (unknown) expectation over $\boldsymbol{\omega}$ with expectation over empirical distribution $\boldsymbol{\omega}_i$. With n observations of $\boldsymbol{\omega}$, or n "scenarios", solve:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

Why is this a good thing to do?

Sample Average Approximation to the Rescue

Let's play same game as in the linear regression case!
Replace (unknown) expectation over $\boldsymbol{\omega}$ with expectation over empirical distribution $\boldsymbol{\omega}_i$. With n observations of $\boldsymbol{\omega}$, or n "scenarios", solve:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

Why is this a good thing to do? Justifications:

- Joint distribution over $\boldsymbol{\omega}$ only exists *in our imagination*, while empirical distribution constructed from data, which is real

Sample Average Approximation to the Rescue

Let's play same game as in the linear regression case!
Replace (unknown) expectation over $\boldsymbol{\omega}$ with expectation over empirical distribution $\boldsymbol{\omega}_i$. With n observations of $\boldsymbol{\omega}$, or n "scenarios", solve:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

Why is this a good thing to do? Justifications:

- Joint distribution over $\boldsymbol{\omega}$ only exists *in our imagination*, while empirical distribution constructed from data, which is real
- As $n \to \infty$, for i.i.d. $\boldsymbol{\omega}^i$, $\hat{\boldsymbol{x}}$ almost surely converges to a minimizer of our two-stage problem under true joint distribution of $\boldsymbol{\omega}$

Sample Average Approximation to the Rescue

Let's play same game as in the linear regression case!
Replace (unknown) expectation over $\boldsymbol{\omega}$ with expectation over empirical distribution $\boldsymbol{\omega}_i$. With n observations of $\boldsymbol{\omega}$, or n "scenarios", solve:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

Why is this a good thing to do? Justifications:

- Joint distribution over $\boldsymbol{\omega}$ only exists *in our imagination*, while empirical distribution constructed from data, which is real
- As $n \to \infty$, for i.i.d. $\boldsymbol{\omega}^i$, $\hat{\boldsymbol{x}}$ almost surely converges to a minimizer of our two-stage problem under true joint distribution of $\boldsymbol{\omega}$
- Who can tell me why we use "arg min" and "a minimizer" here?

Almost Sure Convergence Proof (Sketch)

- Define a sample-average function, redefine expected value

$$\hat{g}_N(\boldsymbol{x}) := \min_{\boldsymbol{y}(\boldsymbol{\omega}^i)} \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i),$$

$$g(\boldsymbol{x}) := \min_{\boldsymbol{y}(\omega)} \mathbb{E}_\omega[\boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega})]$$

Aside

h is the optimal value of a minimization problem. Why is it convex?

Aside

h is the optimal value of a minimization problem. Why is it convex?

$$h(\boldsymbol{x}, \boldsymbol{\omega}) := \min_{\boldsymbol{y}(\boldsymbol{\omega})} \boldsymbol{q}(\boldsymbol{\omega})^\top \boldsymbol{y}(\boldsymbol{\omega}) \text{ s.t. } \boldsymbol{D}(\omega)\boldsymbol{x} + \boldsymbol{F}(\omega)\boldsymbol{y}(\boldsymbol{\omega}) \leq \boldsymbol{d}(\boldsymbol{\omega})$$

Aside

h is the optimal value of a minimization problem. Why is it convex?

$$h(\boldsymbol{x}, \boldsymbol{\omega}) := \min_{\boldsymbol{y}(\boldsymbol{\omega})} \boldsymbol{q}(\boldsymbol{\omega})^\top \boldsymbol{y}(\boldsymbol{\omega}) \text{ s.t. } \boldsymbol{D}(\omega)\boldsymbol{x} + \boldsymbol{F}(\omega)\boldsymbol{y}(\boldsymbol{\omega}) \leq \boldsymbol{d}(\boldsymbol{\omega})$$

Duality!

$$h(\boldsymbol{x}, \boldsymbol{\omega}) = \max_{\boldsymbol{\mu}(\omega)} \quad (\boldsymbol{d}(\boldsymbol{\omega}) - \boldsymbol{D}(\omega)\boldsymbol{x})^\top \boldsymbol{\mu}(\omega) \text{ s.t. } \boldsymbol{F}(\omega)^\top \boldsymbol{\mu}(\omega) = \boldsymbol{q}(\boldsymbol{\omega}), \boldsymbol{\mu}(\omega) \leq \boldsymbol{0}$$

Aside

h is the optimal value of a minimization problem. Why is it convex?

$$h(\boldsymbol{x}, \boldsymbol{\omega}) := \min_{\boldsymbol{y}(\boldsymbol{\omega})} \boldsymbol{q}(\boldsymbol{\omega})^\top \boldsymbol{y}(\boldsymbol{\omega}) \text{ s.t. } \boldsymbol{D}(\omega)\boldsymbol{x} + \boldsymbol{F}(\omega)\boldsymbol{y}(\boldsymbol{\omega}) \leq \boldsymbol{d}(\boldsymbol{\omega})$$

Duality!

$$h(\boldsymbol{x}, \boldsymbol{\omega}) = \max_{\boldsymbol{\mu}(\omega)} \quad (\boldsymbol{d}(\boldsymbol{\omega}) - \boldsymbol{D}(\omega)\boldsymbol{x})^\top \boldsymbol{\mu}(\omega) \text{ s.t. } \boldsymbol{F}(\omega)^\top \boldsymbol{\mu}(\omega) = \boldsymbol{q}(\boldsymbol{\omega}), \boldsymbol{\mu}(\omega) \leq \boldsymbol{0}$$

$h(\boldsymbol{x}, \boldsymbol{\omega})$ is the pointwise maximum of functions linear in $\boldsymbol{x}$, hence convex

Aside

h is the optimal value of a minimization problem. Why is it convex?

$$h(\boldsymbol{x}, \boldsymbol{\omega}) := \min_{\boldsymbol{y}(\boldsymbol{\omega})} \boldsymbol{q}(\boldsymbol{\omega})^\top \boldsymbol{y}(\boldsymbol{\omega}) \text{ s.t. } \boldsymbol{D}(\omega)\boldsymbol{x} + \boldsymbol{F}(\omega)\boldsymbol{y}(\boldsymbol{\omega}) \leq \boldsymbol{d}(\boldsymbol{\omega})$$

Duality!

$$h(\boldsymbol{x}, \boldsymbol{\omega}) = \max_{\boldsymbol{\mu}(\omega)} \quad (\boldsymbol{d}(\boldsymbol{\omega}) - \boldsymbol{D}(\omega)\boldsymbol{x})^\top \boldsymbol{\mu}(\omega) \text{ s.t. } \boldsymbol{F}(\omega)^\top \boldsymbol{\mu}(\omega) = \boldsymbol{q}(\boldsymbol{\omega}), \boldsymbol{\mu}(\omega) \leq \boldsymbol{0}$$

$h(\boldsymbol{x}, \boldsymbol{\omega})$ is the pointwise maximum of functions linear in $\boldsymbol{x}$, hence convex

Pointwise maximum also reveals h is continuous on its domain

Almost Sure Convergence Proof (Sketch)

- Define a sample-average function, redefine expected value

$$\hat{g}_N(\boldsymbol{x}) := \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{x}, \boldsymbol{\omega}^i),$$
$$g(\boldsymbol{x}) := \boldsymbol{c}^\top \boldsymbol{x} + \mathbb{E}_\omega[h(\boldsymbol{x}, \boldsymbol{\omega})]$$

- By SLLN, continuity of g_N, g: $g_N(\boldsymbol{x}) \overset{a.s.}{\to} g(\boldsymbol{x}) \; \forall \boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$

[1] See Corollary 3 of "Monte Carlo Sampling Methods" by Shapiro (2003) for details.

Almost Sure Convergence Proof (Sketch)

- Define a sample-average function, redefine expected value

$$\hat{g}_N(\boldsymbol{x}) := \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{x}, \boldsymbol{\omega}^i),$$
$$g(\boldsymbol{x}) := \boldsymbol{c}^\top \boldsymbol{x} + \mathbb{E}_\omega[h(\boldsymbol{x}, \boldsymbol{\omega})]$$

- By SLLN, continuity of g_N, g: $g_N(\boldsymbol{x}) \overset{a.s.}{\to} g(\boldsymbol{x}) \ \forall \boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$
- Therefore, (under mild conditions[1]), $\inf_{\boldsymbol{x}} g_N(\boldsymbol{x}) \overset{a.s.}{\to} \inf_{\boldsymbol{x}} g(\boldsymbol{x})$

[1] See Corollary 3 of "Monte Carlo Sampling Methods" by Shapiro (2003) for details.

When Things go Wrong, as They Sometimes Will

Let's look at our sample-average approximation again:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

What can go wrong?

When Things go Wrong, as They Sometimes Will

Let's look at our sample-average approximation again:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

What can go wrong? In practice, we have a finite number of observations. That means:

When Things go Wrong, as They Sometimes Will

Let's look at our sample-average approximation again:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

What can go wrong? In practice, we have a finite number of observations. That means:

- $\hat{\boldsymbol{x}}$ may not be feasible for unseen $\boldsymbol{\omega}$'s

When Things go Wrong, as They Sometimes Will

Let's look at our sample-average approximation again:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

What can go wrong? In practice, we have a finite number of observations. That means:

- $\hat{\boldsymbol{x}}$ may not be feasible for unseen $\boldsymbol{\omega}$'s
    - Can include all extreme points of joint dist of ω, or if $h(\boldsymbol{x}, \boldsymbol{\omega}^i)$ is (almost surely) feasible for any $\boldsymbol{x}$—(relatively) complete recourse

When Things go Wrong, as They Sometimes Will

Let's look at our sample-average approximation again:

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \end{aligned}$$

What can go wrong? In practice, we have a finite number of observations. That means:

- $\hat{\boldsymbol{x}}$ may not be feasible for unseen $\boldsymbol{\omega}$'s
  - Can include all extreme points of joint dist of ω, or if $h(\boldsymbol{x}, \boldsymbol{\omega}^i)$ is (almost surely) feasible for any $\boldsymbol{x}$—(relatively) complete recourse
- $\hat{\boldsymbol{x}}_N$ might be far from $\boldsymbol{x}^\star$, especially if N small relative to dim of $\boldsymbol{x}$
  - A motivation for distributionally robust optimization—see later

Let's break for five minutes.

Then talk about how to solve these problems

Sample Average Approximation: Algorithmics

First Strategy: Solve the Deterministic Equivalent

We can view the sample-average approximation as one big linear optimization problem and throw it to Mosek or Gurobi

First Strategy: Solve the Deterministic Equivalent

We can view the sample-average approximation as one big linear optimization problem and throw it to Mosek or Gurobi

- Make a copy of $\boldsymbol{y}^i$ for each scenario $\boldsymbol{\omega}^i$ and solve

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

First Strategy: Solve the Deterministic Equivalent

We can view the sample-average approximation as one big linear optimization problem and throw it to Mosek or Gurobi

- Make a copy of $\boldsymbol{y}^i$ for each scenario $\boldsymbol{\omega}^i$ and solve

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \end{aligned}$$

- Pros: very quick to code, if it works, then we are done

First Strategy: Solve the Deterministic Equivalent

We can view the sample-average approximation as one big linear optimization problem and throw it to Mosek or Gurobi

- Make a copy of $\boldsymbol{y}^i$ for each scenario $\boldsymbol{\omega}^i$ and solve

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

- Pros: very quick to code, if it works, then we are done
- Good first thing to try

First Strategy: Solve the Deterministic Equivalent

We can view the sample-average approximation as one big linear optimization problem and throw it to Mosek or Gurobi

- Make a copy of $\boldsymbol{y}^i$ for each scenario $\boldsymbol{\omega}^i$ and solve

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \end{aligned}$$

- Pros: very quick to code, if it works, then we are done
- Good first thing to try
- Cons: this optimization problem might be big. Really big

First Strategy: Solve the Deterministic Equivalent

We can view the sample-average approximation as one big linear optimization problem and throw it to Mosek or Gurobi

- Make a copy of $\boldsymbol{y}^i$ for each scenario $\boldsymbol{\omega}^i$ and solve

$$
\begin{aligned}
\hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}
\end{aligned}
$$

- Pros: very quick to code, if it works, then we are done
- Good first thing to try
- Cons: this optimization problem might be big. Really big
- Example: electricity market with random demand at 20 nodes that can independently be "low" or "high"

First Strategy: Solve the Deterministic Equivalent

We can view the sample-average approximation as one big linear optimization problem and throw it to Mosek or Gurobi

- Make a copy of $\boldsymbol{y}^i$ for each scenario $\boldsymbol{\omega}^i$ and solve

$$\begin{aligned} \hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \end{aligned}$$

- Pros: very quick to code, if it works, then we are done
- Good first thing to try
- Cons: this optimization problem might be big. Really big
- Example: electricity market with random demand at 20 nodes that can independently be "low" or "high" That's $2^{20} = 1048576$ copies of $\boldsymbol{y}$, which is intractable for a real market

First Strategy: Solve the Deterministic Equivalent

We can view the sample-average approximation as one big linear optimization problem and throw it to Mosek or Gurobi

- Make a copy of $\boldsymbol{y}^i$ for each scenario $\boldsymbol{\omega}^i$ and solve

$$
\begin{aligned}
\hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}
\end{aligned}
$$

- Pros: very quick to code, if it works, then we are done
- Good first thing to try
- Cons: this optimization problem might be big. Really big
- Example: electricity market with random demand at 20 nodes that can independently be "low" or "high" That's $2^{20} = 1048576$ copies of $\boldsymbol{y}$, which is intractable for a real market
- Still, you can sometimes do well by subsampling the scenarios (Shapiro and Homem-de-Mello, 1998)

Second Strategy: Decompose the Problem

What optimizers usually do: use a decomposition scheme called Benders decomposition (sometimes called the “L-shaped” method)

Second Strategy: Decompose the Problem

What optimizers usually do: use a decomposition scheme called Benders decomposition (sometimes called the "L-shaped" method)

Consider

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b} \end{aligned}$$

Let $\theta \ge \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$ be an epigraph variable

Second Strategy: Decompose the Problem

$$
\begin{aligned}
\min_{\boldsymbol{x} \in \mathbb{R}^n, \theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}.
\end{aligned}
$$

Second Strategy: Decompose the Problem

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n, \theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

(Sketch) We iteratively

- Solve this “master” problem to find an optimal $\boldsymbol{x}$

Second Strategy: Decompose the Problem

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n, \theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

(Sketch) We iteratively

- Solve this "master" problem to find an optimal $\boldsymbol{x}$
- Evaluate $1/n \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$ and add inequalities which model
  - $\theta \geq \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$
  - For $\boldsymbol{x}$ to be feasible, there is a feasible $\boldsymbol{y}(\boldsymbol{\omega}^i)$ in each scenario $\boldsymbol{\omega}^i$

until we converge.

Second Strategy: Decompose the Problem

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n, \theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

(Sketch) We iteratively

- Solve this "master" problem to find an optimal $\boldsymbol{x}$
- Evaluate $1/n \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$ and add inequalities which model
    - $\theta \geq \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$
    - For $\boldsymbol{x}$ to be feasible, there is a feasible $\boldsymbol{y}(\boldsymbol{\omega}^i)$ in each scenario $\boldsymbol{\omega}^i$

until we converge. We never model $\boldsymbol{y}(\boldsymbol{\omega}^i)$, so we replaced one intractable problem with a sequence of (possibly many) tractable ones

Second Strategy: Decompose the Problem

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n, \theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

(Sketch) We iteratively

- Solve this "master" problem to find an optimal $\boldsymbol{x}$
- Evaluate $1/n \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$ and add inequalities which model
  - $\theta \geq \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$
  - For $\boldsymbol{x}$ to be feasible, there is a feasible $\boldsymbol{y}(\boldsymbol{\omega}^i)$ in each scenario $\boldsymbol{\omega}^i$

until we converge. We never model $\boldsymbol{y}(\boldsymbol{\omega}^i)$, so we replaced one intractable problem with a sequence of (possibly many) tractable ones

Remark: About to go through how this works in gory detail. However, I find the best way to understand this method is to code it for yourself.

Benders Decomposition

Suppose we solve

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n, \theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

and obtain some solution $\boldsymbol{x}$. Two cases:

- There is some scenario ω^i for which no $\boldsymbol{y}(\boldsymbol{\omega})$ can make the scenario feasible $\rightarrow$ we need to tell the master problem that this $\boldsymbol{x}$ is infeasible, via a *feasibility cut*

Benders Decomposition

Suppose we solve

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n, \theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

and obtain some solution $\boldsymbol{x}$. Two cases:

- There is some scenario ω^i for which no $\boldsymbol{y}(\boldsymbol{\omega})$ can make the scenario feasible $\rightarrow$ we need to tell the master problem that this $\boldsymbol{x}$ is infeasible, via a *feasibility cut*
- Every scenario ω^i is feasible $\rightarrow$ we need to tell the master problem how much $\boldsymbol{x}$ costs via an *optimality cut*

Benders Decomposition: Feasibility Cut

Suppose we solve

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n, \theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}. \end{aligned}$$

and obtain some solution $\boldsymbol{x}$ such that in scenario i no $\boldsymbol{y}(\boldsymbol{\omega})$ can make the scenario feasible.

Benders Decomposition: Feasibility Cut

Suppose we solve

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^n, \theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}. \end{aligned}$$

and obtain some solution $\boldsymbol{x}$ such that in scenario i no $\boldsymbol{y}(\boldsymbol{\omega})$ can make the scenario feasible. Then, the dual problem in this scenario is unbounded (why?), so there is some $\boldsymbol{\mu}(\boldsymbol{\omega}^i)$ such that

$$(\boldsymbol{d}(\boldsymbol{\omega}) - \boldsymbol{D}(\omega)\boldsymbol{x})^\top \boldsymbol{\mu}(\omega) > 0, \;\; \boldsymbol{F}(\omega)^\top \boldsymbol{\mu}(\omega) = \mathbf{0}, \boldsymbol{\mu}(\omega) \leq \mathbf{0}.$$

Benders Decomposition: Feasibility Cut

Suppose we solve

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n,\theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}. \end{aligned}$$

and obtain some solution $\boldsymbol{x}$ such that in scenario i no $\boldsymbol{y}(\boldsymbol{\omega})$ can make the scenario feasible. Then, the dual problem in this scenario is unbounded (why?), so there is some $\boldsymbol{\mu}(\boldsymbol{\omega}^i)$ such that

$$(\boldsymbol{d}(\boldsymbol{\omega}) - \boldsymbol{D}(\omega)\boldsymbol{x})^\top \boldsymbol{\mu}(\omega) > 0,\ \boldsymbol{F}(\omega)^\top \boldsymbol{\mu}(\omega) = \mathbf{0}, \boldsymbol{\mu}(\omega) \le \mathbf{0}.$$

Therefore, we fix $\boldsymbol{\mu}(\boldsymbol{\omega}^i)$ and impose the feasibility cut

$$(\boldsymbol{d}(\boldsymbol{\omega}^i) - \boldsymbol{D}(\boldsymbol{\omega}^i)\boldsymbol{x})^\top \boldsymbol{\mu}(\boldsymbol{\omega}^i) \le 0,$$

in the master problem, where everything but $\boldsymbol{x}$ is data

In This Case, The Master Problem Now Looks Like

$$
\begin{aligned}
\min_{\boldsymbol{x}\in\mathbb{R}^n,\theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \\
& (\boldsymbol{d}(\omega^i) - \boldsymbol{D}(\omega^i)\boldsymbol{x})^\top \boldsymbol{\mu}(\omega^i) \leq 0.
\end{aligned}
$$

Case Two: Each Scenario Was Feasible

θ usually underestimates $1/n \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$

Case Two: Each Scenario Was Feasible

θ usually underestimates $1/n \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$

Need cut involving θ, which tells master problem what $\boldsymbol{x}$ costs

Case Two: Each Scenario Was Feasible

θ usually underestimates $1/n \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$

Need cut involving θ, which tells master problem what $\boldsymbol{x}$ costs

By strong duality

$$\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) = 1/n \sum_{i=1}^{n} (\boldsymbol{d}(\boldsymbol{\omega}^i) - \boldsymbol{D}(\boldsymbol{\omega}^i)\boldsymbol{x})^\top \boldsymbol{\mu}(\boldsymbol{\omega}^i),$$

where $\boldsymbol{\mu}(\boldsymbol{\omega}^i)$, dual-optimal in scenario i, is data

Case Two: Each Scenario Was Feasible

θ usually underestimates $1/n \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$

Need cut involving θ, which tells master problem what $\boldsymbol{x}$ costs

By strong duality

$$\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) = 1/n \sum_{i=1}^{n} (\boldsymbol{d}(\boldsymbol{\omega}^i) - \boldsymbol{D}(\boldsymbol{\omega}^i)\boldsymbol{x})^\top \boldsymbol{\mu}(\boldsymbol{\omega}^i),$$

where $\boldsymbol{\mu}(\boldsymbol{\omega}^i)$, dual-optimal in scenario i, is data

By weak duality, for any $\bar{\boldsymbol{x}}$

$$\frac{1}{n}\sum_{i=1}^{n} h(\bar{\boldsymbol{x}}, \boldsymbol{\omega}^i) \geq \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{d}(\boldsymbol{\omega}^i) - \boldsymbol{D}(\boldsymbol{\omega}^i)\bar{\boldsymbol{x}})^\top \boldsymbol{\mu}(\boldsymbol{\omega}^i),$$

where everything but $\bar{\boldsymbol{x}}$ is data

Case Two: Each Scenario Was Feasible

θ usually underestimates $1/n \sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i)$

Need cut involving θ, which tells master problem what $\boldsymbol{x}$ costs

By strong duality

$$\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}, \boldsymbol{\omega}^i) = 1/n \sum_{i=1}^{n} (\boldsymbol{d}(\boldsymbol{\omega}^i) - \boldsymbol{D}(\boldsymbol{\omega}^i)\boldsymbol{x})^\top \boldsymbol{\mu}(\boldsymbol{\omega}^i),$$

where $\boldsymbol{\mu}(\boldsymbol{\omega}^i)$, dual-optimal in scenario i, is data

By weak duality, for any $\bar{\boldsymbol{x}}$

$$\frac{1}{n}\sum_{i=1}^{n} h(\bar{\boldsymbol{x}}, \boldsymbol{\omega}^i) \geq \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{d}(\boldsymbol{\omega}^i) - \boldsymbol{D}(\boldsymbol{\omega}^i)\bar{\boldsymbol{x}})^\top \boldsymbol{\mu}(\boldsymbol{\omega}^i),$$

where everything but $\bar{\boldsymbol{x}}$ is data

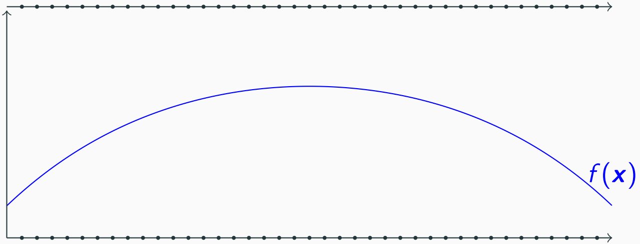
Therefore, we add cut

$$\theta \geq \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{d}(\boldsymbol{\omega}^i) - \boldsymbol{D}(\boldsymbol{\omega}^i)\boldsymbol{x})^\top \boldsymbol{\mu}(\boldsymbol{\omega}^i)$$

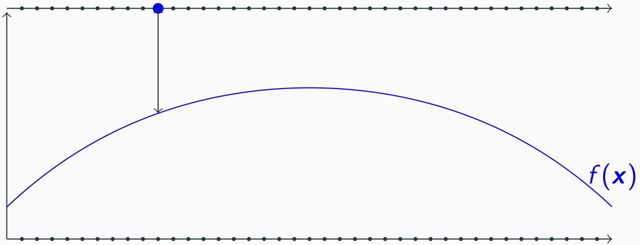
The Master Problem Might Now Look Like

$$
\begin{aligned}
\min_{\boldsymbol{x}\in\mathbb{R}^n,\theta} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \theta \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \\
& \theta \geq \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{d}(\boldsymbol{\omega}^i) - \boldsymbol{D}(\boldsymbol{\omega}^i)\boldsymbol{x})^\top \boldsymbol{\mu}(\boldsymbol{\omega}^i), \\
& (\boldsymbol{d}(\boldsymbol{\omega}^i) - \boldsymbol{D}(\boldsymbol{\omega}^i)\boldsymbol{x})^\top \boldsymbol{\mu}(\boldsymbol{\omega}^i) \leq 0.
\end{aligned}
$$

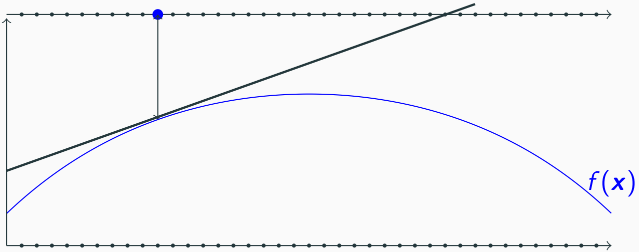
Benders Decomposition, in 1000 words



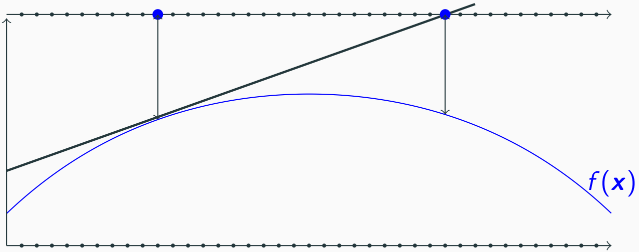
Benders Decomposition, in 1000 words



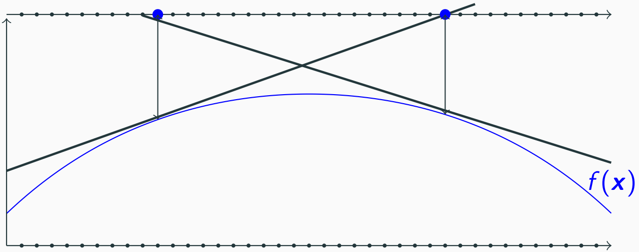
Benders Decomposition, in 1000 words



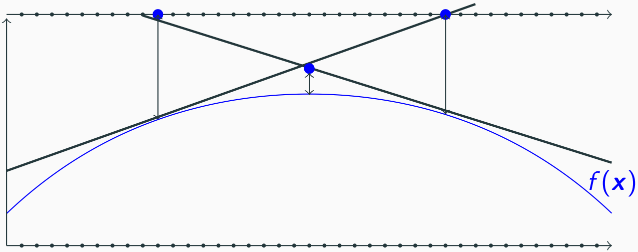
Benders Decomposition, in 1000 words



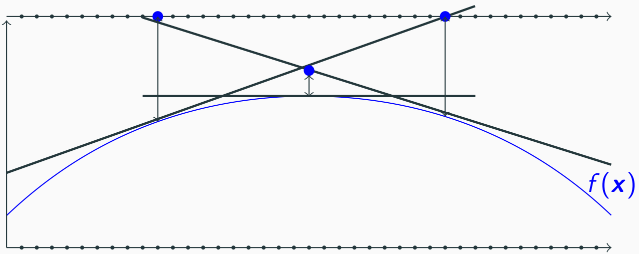
Benders Decomposition, in 1000 words



Benders Decomposition, in 1000 words



Benders Decomposition, in 1000 words



Sample Average Approximation: Code
You will write this yourself in the first assignment :-)

Can we do Better? Ridge Regression and Sample-Average Approximation

Can we do Better Than the Sample-Average Approximation?

Returning to Linear Regression

Statisticians don't solve problems like

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{n} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2$$

to pick $\boldsymbol{\beta}$, despite SAA's properties. Why not?

Returning to Linear Regression

Statisticians don't solve problems like

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{n} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2$$

to pick $\boldsymbol{\beta}$, despite SAA's properties. Why not?

Because n is finite; we want $\boldsymbol{\beta}$ to perform as well as possible on an unseen observation $(\boldsymbol{x}_i, y_i)$, not just minimize training error.

Returning to Linear Regression

Statisticians don't solve problems like

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2$$

to pick $\boldsymbol{\beta}$, despite SAA's properties. Why not?

Because n is finite; we want $\boldsymbol{\beta}$ to perform as well as possible on an unseen observation $(\boldsymbol{x}_i, y_i)$, not just minimize training error.They solve

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + R(\boldsymbol{\beta}),$$

where $R(\cdot)$ is a regularization term, e.g., $\frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$ for appropriately chosen λ, γ (elastic net method, Zou and Hastie 2005).

Returning to Linear Regression

Statisticians don't solve problems like

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{n} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2$$

to pick $\boldsymbol{\beta}$, despite SAA's properties. Why not?

Because n is finite; we want $\boldsymbol{\beta}$ to perform as well as possible on an unseen observation $(\boldsymbol{x}_i, y_i)$, not just minimize training error. They solve

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{n} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + R(\boldsymbol{\beta}),$$

where $R(\cdot)$ is a regularization term, e.g., $\frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$ for appropriately chosen λ, γ (elastic net method, Zou and Hastie 2005).

This usually performs better out-of-sample.

How is This Related to SAA?

- In the 2000s, sample-average approximation was a very popular method for optimizing under uncertainty

How is This Related to SAA?

- In the 2000s, sample-average approximation was a very popular method for optimizing under uncertainty
- In the early 2010s, the community became more aware of the danger of overfitting, and in particular that SAA tends to be overly optimistic w.r.t. its out-of-sample performance. Since then, variants of SAA that account for overfitting with better finite-sample guarantees have become popular

How is This Related to SAA?

- In the 2000s, sample-average approximation was a very popular method for optimizing under uncertainty
- In the early 2010s, the community became more aware of the danger of overfitting, and in particular that SAA tends to be overly optimistic w.r.t. its out-of-sample performance. Since then, variants of SAA that account for overfitting with better finite-sample guarantees have become popular
- We still teach SAA, because you need to understand SAA first

How is This Related to SAA?

- In the 2000s, sample-average approximation was a very popular method for optimizing under uncertainty
- In the early 2010s, the community became more aware of the danger of overfitting, and in particular that SAA tends to be overly optimistic w.r.t. its out-of-sample performance. Since then, variants of SAA that account for overfitting with better finite-sample guarantees have become popular
- We still teach SAA, because you need to understand SAA first
- Variants intimately related to distributional robustness (see Lecture 8)

How is This Related to SAA?

- In the 2000s, sample-average approximation was a very popular method for optimizing under uncertainty
- In the early 2010s, the community became more aware of the danger of overfitting, and in particular that SAA tends to be overly optimistic w.r.t. its out-of-sample performance. Since then, variants of SAA that account for overfitting with better finite-sample guarantees have become popular
- We still teach SAA, because you need to understand SAA first
- Variants intimately related to distributional robustness (see Lecture 8)
- For more on this, see: Bertsimas, Dimitris, Vishal Gupta, and Nathan Kallus. "Robust sample average approximation." Mathematical Programming 171.1 (2018): 217-282.

Extension: Benders Decomposition for Facility Location

See slides by Fischetti (2017)

Suggested Readings

Suggested Readings to Accompany Today's Lecture

A friendly reminder:

> *"To get as much out of this class as possible, we suggest that you spend at least as much time on reading the papers and textbooks referenced in the lectures/reviewing the lectures as you spend in class." — The syllabus*

Recommended reading:

- Shapiro, Dentcheva, Ruszczynski *Lectures on Stochastic Programming: Modeling and Theory* (2013), Chapters 1.1 and 2.

Optional further reading:

- Recht *Lecture* 1. In CS294 The Mathematics of Data Science lecture notes, UC Berkeley (2013).
- Kim, Pasupathy, Henderson *A Guide to Sample-Average Approximation*. In: Handbook of simulation optimization (2015).

Thank you, and see you next week!