# STOCHASTIC OPTIMIZATION

# AN EP VIEWPOINT

SP 23, STUDY GR.

# EXAMPLE I (MLE)

$$Y \sim p_{\theta_0}(x), \quad x \in \mathcal{X}.$$

$p_{\theta_0}$ is a density w.r.t

a $\sigma$-finite measure.

$\theta_0 \in \Theta$ is unknown

Available data: $Y_1, Y_2, \ldots, Y_n$ iid

## EXAMPLE I (MLE) contd...

Notice that $\forall \theta \in \Theta$

$$\mathbb{E} \log p_{\theta_0}(Y) \geq \mathbb{E} \log p_{\theta}(Y) \quad \text{---(1)}$$

Why?

## EXAMPLE I (MLE) contd...

(1) implies that

$$\theta_0 \in \operatorname{argmin} \mathbb{E} \log p_\theta(Y)$$

## EXAMPLE I (MLE) contd...

How to estimate $\theta_0$?

Suppose $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} P_{\theta_0}$

Solve:

$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\arg\min} \cdot \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(Y_i)$$

→ max. likelihood estimator.

## EXAMPLE I (MLE) contd...

The alternate EP viewpoint.

$$\log P_\theta(Y) \equiv F(\theta, Y)$$

where $F : \Theta \times \mathcal{Y} \to \mathbb{R}$

## EXAMPLE I (MLE) contd...

The alternate EP viewpoint.

$$\theta_0 \in \arg\min \mathbb{E}\left[F(\theta, Y)\right]$$

$$\hat{\theta}_n \in \arg\min \frac{1}{n} \sum_{i=1}^{n} F(\theta, Y_i)$$

# EXAMPLE I (MLE) contd...

Main Questions:

(Q.1) Does $\hat{\theta}_n \longrightarrow \theta_0$ in some sense?

(Q.2) Is there a "CLT" on $\hat{\theta}_n$?

(Q.3) What if $\hat{\theta}_n$ can be solved only approximately?

# EXAMPLE I (MLE) contd...

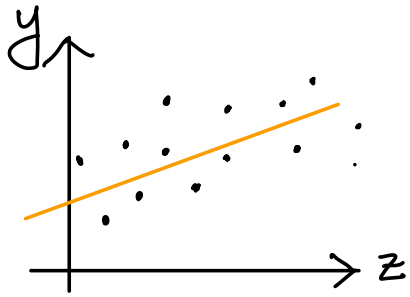
Food for Thought.

(a) Construct an example where $\Theta$ is finite dimensional.

(b) Construct an example where $\Theta$ is infinite dimensional.

# EXAMPLE II (Least Squares)

$(\tilde{Y}, Z)$




"response" $\tilde{Y}$ ; "covariate" $Z$

$\tilde{Y} \in \mathbb{R}$ ; $Z \in \mathcal{L}$.

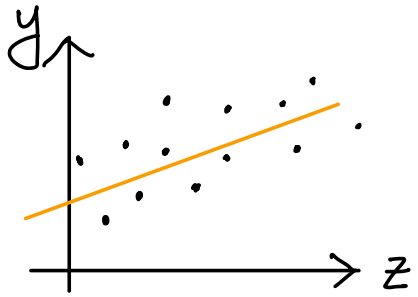
Model:

$$\tilde{Y} = \theta_0(Z) + W$$

$\theta_0$: "true reg. fn."

$W$: "error"

# Least Squares ...




Observe data

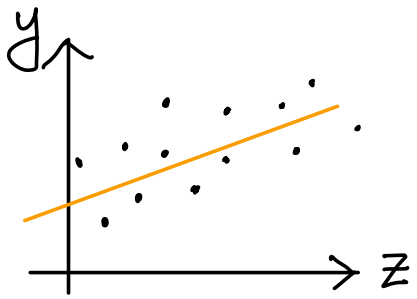
$$(\tilde{Y}_i, Z_i), \quad i = 1, 2, \ldots, n$$

- $(\tilde{Y}_i, Z_i)$ iid.
- $W_1, W_2, \ldots$ independent with $\mathbb{E}[W_i] = 0$, $\mathrm{Var}(W_i) \leq \sigma_0^2 < \infty$.
- $\theta_0 : \mathcal{L} \to \mathbb{R}$; $\theta_0 \in \Theta$

↓

"regression class"

# Least Squares ...




$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left( \theta(Z_i) - \tilde{Y}_i \right)^2$$

"least squares estimator"

## EXAMPLE I (MLE) contd...

The alternate EP viewpoint.

$$\left(\theta(Z) - \tilde{Y}\right)^2 \equiv F(\theta, Y)$$

Where $F : \Theta \times \mathcal{Y} \longrightarrow \mathbb{R}$

$$\theta_0 \in \arg\min \mathbb{E}\left[F(\theta, Y)\right]$$

$$\hat{\theta}_n \in \arg\min \frac{1}{n} \sum_{i=1}^{n} F(\theta, Y_i)$$

## Least Squares ...

(Q.1) Does $\hat{\theta}_n \longrightarrow \theta_0$ in some sense?

(Q.2) Is there a "CLT" on $\hat{\theta}_n$?

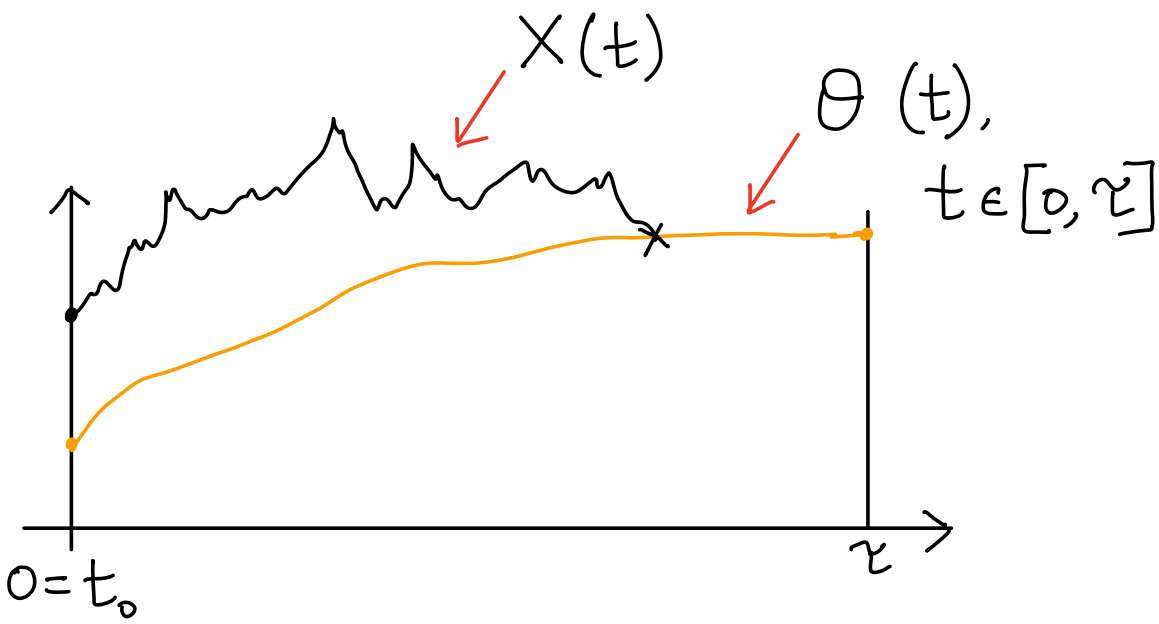
(Q.3) What if $\hat{\theta}_n$ can be solved only approximately?

Least Squares ...

Food for thought.

Linear regression, logistic regression, isotonic regression, deep learning fall in this framework. Convince yourself...

# EXAMPLE III (Limit Order Book)




$\{X(t),\ t \in [0, \infty)\}$ "asset price process"

$\theta(\cdot)$ : "execution boundary"

$[0, \tau]$ : "horizon"

LOB ...

$$T := \inf\{t : \theta(t) \geq X(t)\}$$

Purchase price:

$$C^*(\theta) := \begin{cases} \theta(T) & T \leq \tau \\ m(\tau) & T > \tau \end{cases}$$

"market price"

LOB ...

$$\min \; \mathbb{E}\left[ C^*(\theta) \right]$$

$$\text{s.t.} \quad \text{Var}\left(C^*(\theta)\right) \leq \gamma.$$

$$\theta \in \Theta$$

notation.

$$\hat{\theta}_n \in \arg\min \frac{1}{n} \sum_{j=1}^{n} C^*(\theta, Y_j)$$

$$\text{s.t} \quad \frac{1}{n} \sum_{j=1}^{n} \left( C^*(\theta, Y_j) - \bar{C}^*(\theta) \right)^2$$

$$\theta \in \Theta.$$

# LOB ...

(Q.1) Does $\hat{\theta}_n \longrightarrow \theta_0$ in some sense?

(Q.2) Is there a "CLT" on $\hat{\theta}_n$?

(Q.3) What if $\hat{\theta}_n$ can be solved only approximately?

# Stochastic Optimization (unconstrained)

$$\min. \; f_0(\theta) = \mathbb{E}\left[ F(\theta, Y) \right]$$

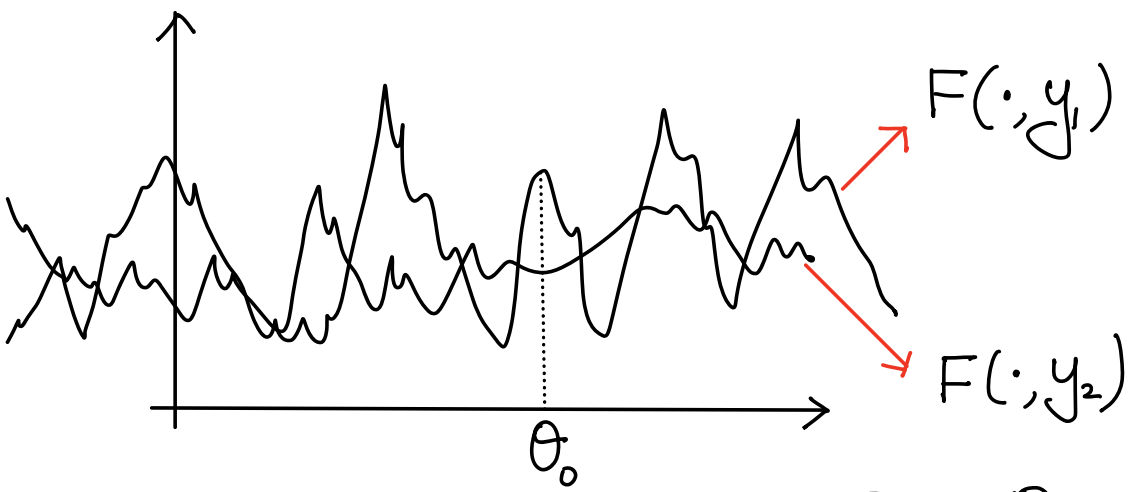
$$\text{s.t.} \quad \theta \in \Theta .$$

$$f_0 : \Theta \to \mathbb{R}, \qquad F : \Theta \times \mathcal{Y} \to \mathbb{R} .$$

$F$ and possibly some of its "derivatives" are observable.

"Function Class" $\mathcal{F}$.

This is the class of "random variables" labeled by $\theta$.




$\Theta \equiv \mathbb{R}$

$$\mathcal{F} := \{ F(\theta, Y),\ \theta \in \Theta \}.$$

We will see that the "richness" of $\mathcal{F}$ will play a very important role.

Recall Stochastic Optimization (unconstrained)

$$\min. \; f(\theta) = \mathbb{E}\left[ F(\theta, Y) \right]$$

$$\text{s.t.} \quad \theta \in \Theta .$$

$$f : \Theta \to \mathbb{R}, \qquad F : \Theta \times \mathcal{Y} \to \mathbb{R}$$

$F$ and possibly some of its "derivatives" are observable.

# M-Estimation

$$\min \frac{1}{n}\sum_{j=1}^{n} F(\theta, Y_j) \qquad \min \mathbb{E}\left[F(\theta, Y)\right]$$

$$\text{s.t.} \quad \theta \in \Theta \qquad \text{s.t} \quad \theta \in \Theta$$

# (Z-Estimation)

solve:

$$\frac{1}{n}\sum_{j=1}^{n} G_1(\theta, Y_j) = 0$$

solve:

$$\mathbb{E}\left[G_1(\theta, Y)\right] = 0$$

A solution $\hat{\theta}_n$ to the M-Estimation problem is an M-Estimator.

Many, many statistical estimators are M-Estimators.

Let's look at examples.

# M-Estimation Example I

Recall MLE.

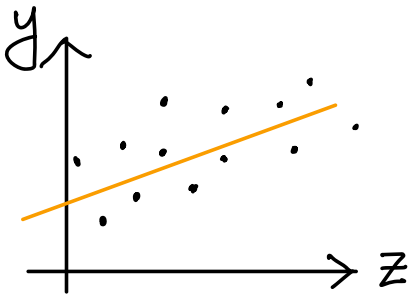
$$X_1, X_2, \dots X_n \overset{iid}{\sim} P_{\theta^*}.$$

$$\min_{\theta \in \Theta} \; -\frac{1}{n} \sum_{j=1}^{n} \log P_\theta(X_j)$$

$$F(\theta, Y_j)$$

# M-Estimation Example II

Regression:




Data: $(X_i, R_i)$, $i = 1, 2, \ldots, n$

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{j=1}^{n} \underbrace{\left(\theta(X_i) - R_i\right)^2}_{F(\theta, Y_i)}$$

"space of reg. functions"

# M-Estimation Example III

Deep Learning

# Z-Estimation Example IV

Mean Estimation.

Set $G_1(\theta, Y) \equiv Y - \theta$

solve:

$$\frac{1}{n}\sum_{j=1}^{n}(Y_j - \theta) = 0$$

$$\Rightarrow \hat{\theta}_n = \bar{Y}$$

$$\theta^* = ?$$

# Z-Estimation Example V

Mean-variance estimation

set $G_1(\theta, Y) = \begin{pmatrix} Y - \theta_1 \\ (Y-\theta_1)^2 - \theta_2 \end{pmatrix}$

solve:

$$\frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} Y_j - \theta_1 \\ (Y_j-\theta_1)^2 - \theta_2 \end{pmatrix} = 0$$

$$\Rightarrow \hat{\theta}_n = \begin{pmatrix} \bar{Y} \\ \hat{\sigma}_n^2 \end{pmatrix} \quad \theta^* = ?$$

# Z-Estimation Example VI

## Ratio Estimation.

set $G_1(\theta, Y) \equiv Y_1 - \theta Y_2$

solve:

$$\frac{1}{n}\sum_{j=1}^{n}\left(Y_{1,j} - \theta Y_{2,j}\right) = 0$$

$$\Rightarrow \hat{\theta}_n = \bar{Y}_1 / \bar{Y}_2 .$$

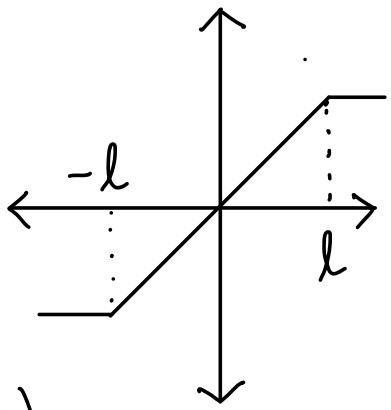
$$\theta^* = ? \quad \theta^* = \frac{E[Y_1]}{E[Y_2]}$$

# Z-Estimation Example VII

## Robust Estimation (Huber, 1964)

set

$$h_l(x) = \begin{cases} -l & x < -l \\ x & |x| \le l \\ l & x > l \end{cases}$$




$$G_1(\theta, Y) \equiv h(Y - \theta)$$

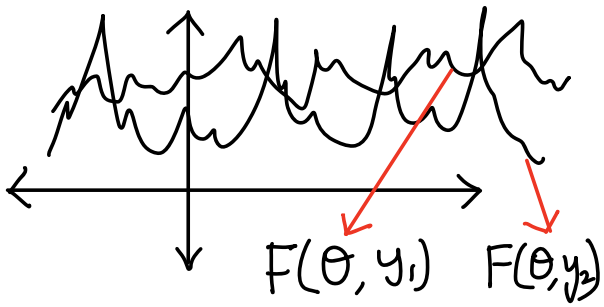
solve:

$$\frac{1}{n} \sum_{j=1}^{n} G_1(\theta, Y_j) = 0$$

( $\theta^*$ satisfies $\mathbb{E}[h'(Y-\theta)] = 0$ )

# So What? (Heuristic Preview)

So What?




EP formalizes and vastly generalizes by looking at

$$\{ F(\theta, Y),\ \theta \in \Theta \}$$

(or $\{ G_1(\theta, Y),\ \theta \in \Theta \}$).

Let's begin our formal study...

# NOTATION

Suppose $Y_1, Y_2, \ldots$ are independent copies of $Y$ having distribution $P$ on $(\mathcal{X}, \mathcal{A})$.

$P_n \longrightarrow$ "empirical distbn."

$$\left( P_n(A) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}_A(Y_j) \right)$$

$$P_n F_\theta \equiv P_n F(\theta, \cdot) = \frac{1}{n} \sum_{j=1}^{n} F(\theta, Y_j) =: \int F_\theta \, dP_n .$$

"expectation of $F(\theta, \cdot)$ under the empirical measure"

Similarly,

$$PF_\theta \equiv PF(\theta, \cdot)$$

$$= \mathbb{E}\left[F(\theta, Y)\right].$$

$$\mathcal{F} := \{ F(\theta, Y),\ \theta \in \Theta \}$$

class of measurable functions labeled by $\theta$.

## Uniform Norm

$$\left\| P_n - P \right\|_{\mathcal{F}} =$$

$$\sup_{\theta \in \Theta} \left| P_n F_\theta - P F_\theta \right|$$

Uniform Law of Large Numbers (ULLN) w.r.t $\mathcal{F}$.

$$\lim_{n\to\infty} \left\| P_n - P \right\|_{\mathcal{F}} = 0 \quad P \text{ a.s.}$$

$\mathcal{F}$ will be called the Glivenko-Cantelli class.

We will write:

$$\int F_\theta \, d(P_n - P)$$

to mean

$$P_n F_\theta - P F_\theta$$

$$= \frac{1}{n} \sum_{j=1}^{n} F(\theta, Y_j) - \mathbb{E}\left[F(\theta, \bar{Y})\right]$$

$$\left\{ \sqrt{n}\left(P_n F_\theta - P F_\theta\right),\ \theta \in \Theta \right\}$$

"empirical process"