Lecture 1: Introduction to zeroth-order optimization (ZOO)

Motivation

I DID
NOTHING
TODAY AND
STILL GOT PAID

Application I: Service System

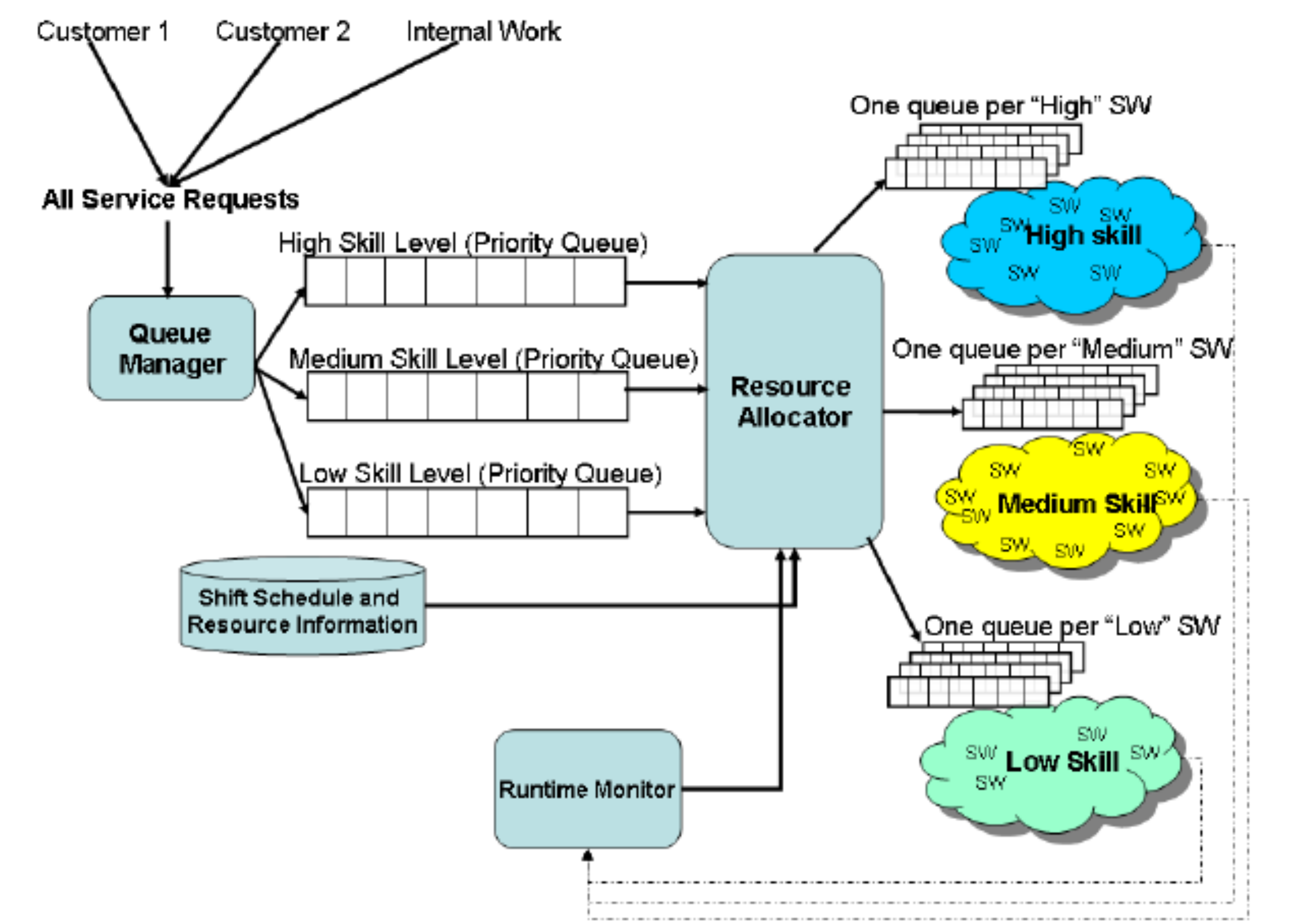


Table 1: Workers $W_{i,j}$

	Skill levels		
Shift	High	Med	Low
S1	1	3	7
S2	0	5	2
S3	3	1	2

Table 2: SLA targets $\gamma_{i,j}$

	Customers	
Priority	Bossy Corp	Cool Inc
P_1	4h	5h
P_2	8h	12h
P_3	24h	48h
P_4	18h	144h

$$x \rightarrow \boxed{\text{Gradient Oracle}} \rightarrow \overset{G \equiv}{\nabla f(x) + \epsilon} \sim N(0,1)$$

$$x_{k+1} = x_k - a_k G_k$$

$$x^* \in \arg\min_x f(x)$$

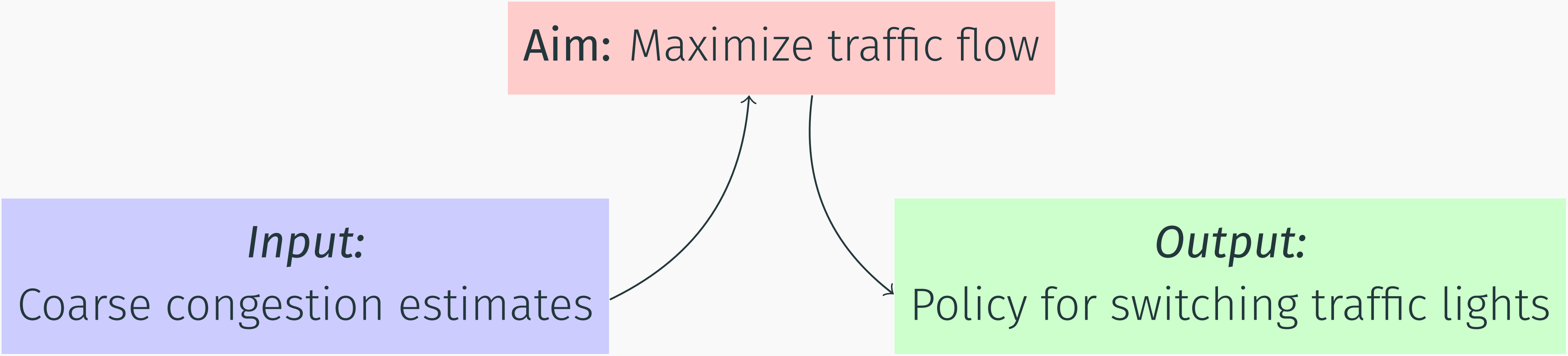
Aim: Find the optimal number of workers for each shift and of each skill level

- that minimizes the labor cost and
- satisfies SLA requirements

Application II: Transportation

On a good day, the traffic is . . .

And on a bad day, it can be . . .



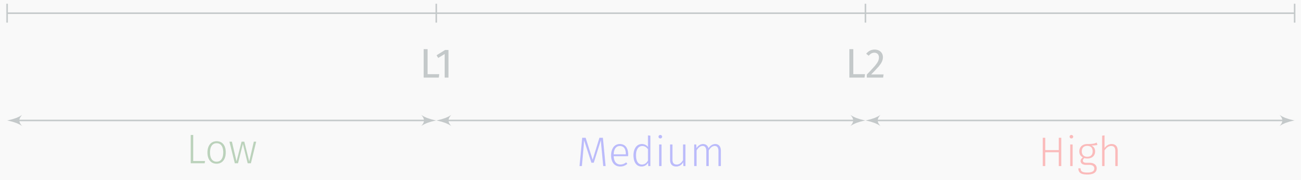
Aim: Maximize traffic flow

Input:
Coarse congestion estimates

Output:
Policy for switching traffic lights

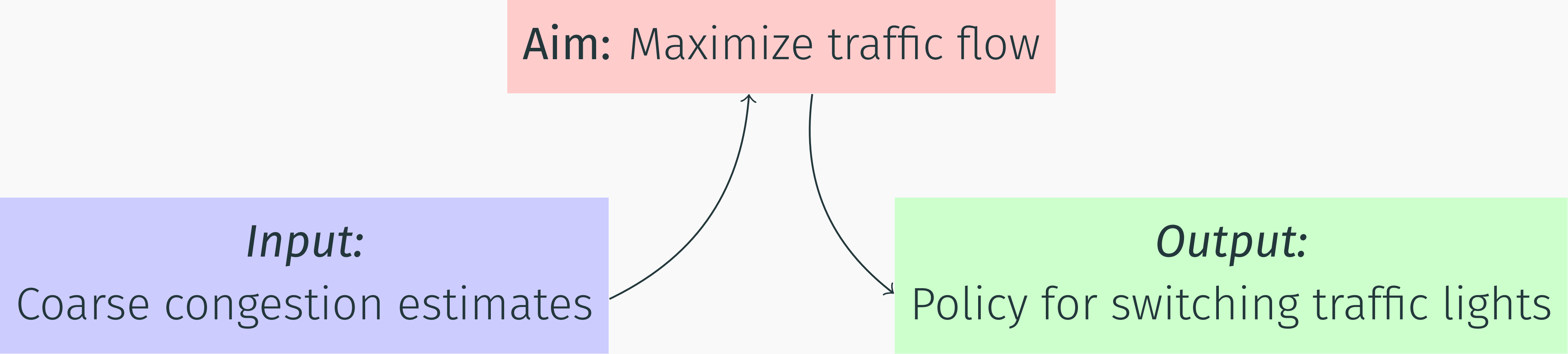
Input: Coarse congestion estimates

Sensor loops at two points along the road



How to switch traffic lights given L1 and L2?

How to choose L1 and L2 for a given policy and road network?



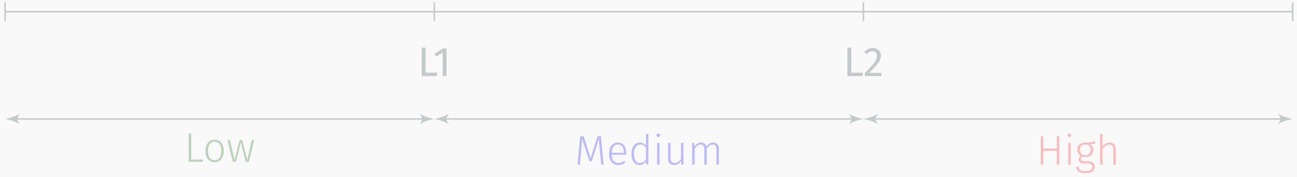
Aim: Maximize traffic flow

Input:
Coarse congestion estimates

Output:
Policy for switching traffic lights

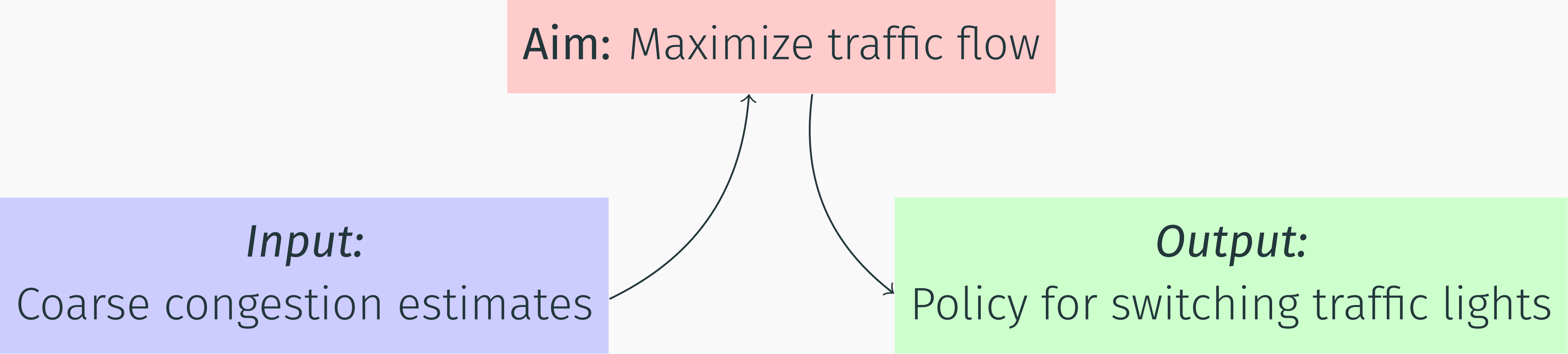
Input: Coarse congestion estimates

Sensor loops at two points along the road



How to switch traffic lights given L1 and L2?

How to choose L1 and L2 for a given policy and road network?



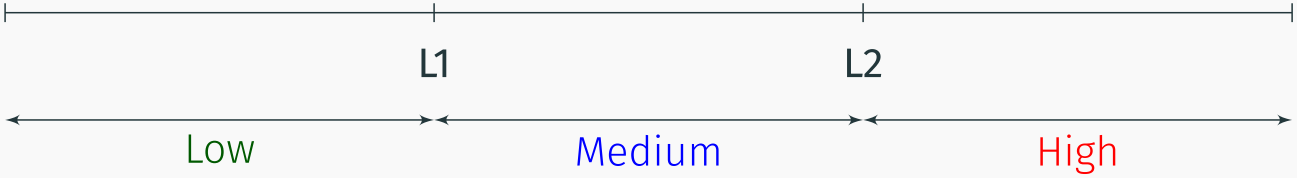
Aim: Maximize traffic flow

Input:
Coarse congestion estimates

Output:
Policy for switching traffic lights

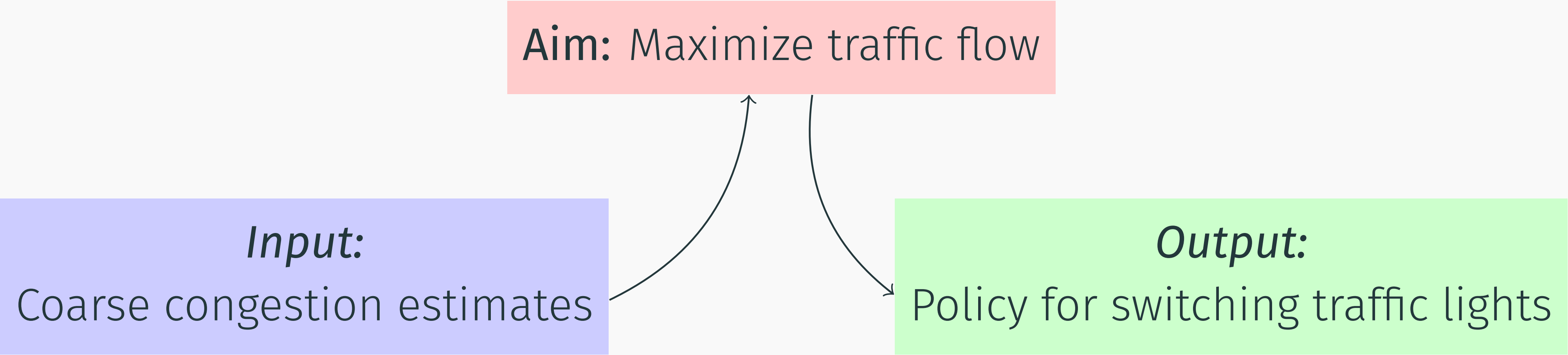
Input: Coarse congestion estimates

Sensor loops at two points along the road



How to switch traffic lights given L1 and L2?

How to choose L1 and L2 for a given policy and road network?



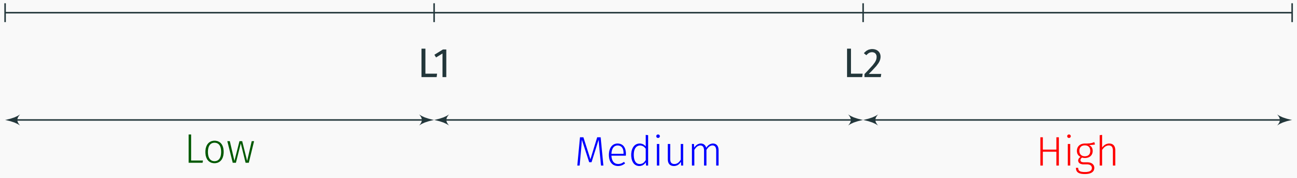
Aim: Maximize traffic flow

Input:
Coarse congestion estimates

Output:
Policy for switching traffic lights

Input: Coarse congestion estimates

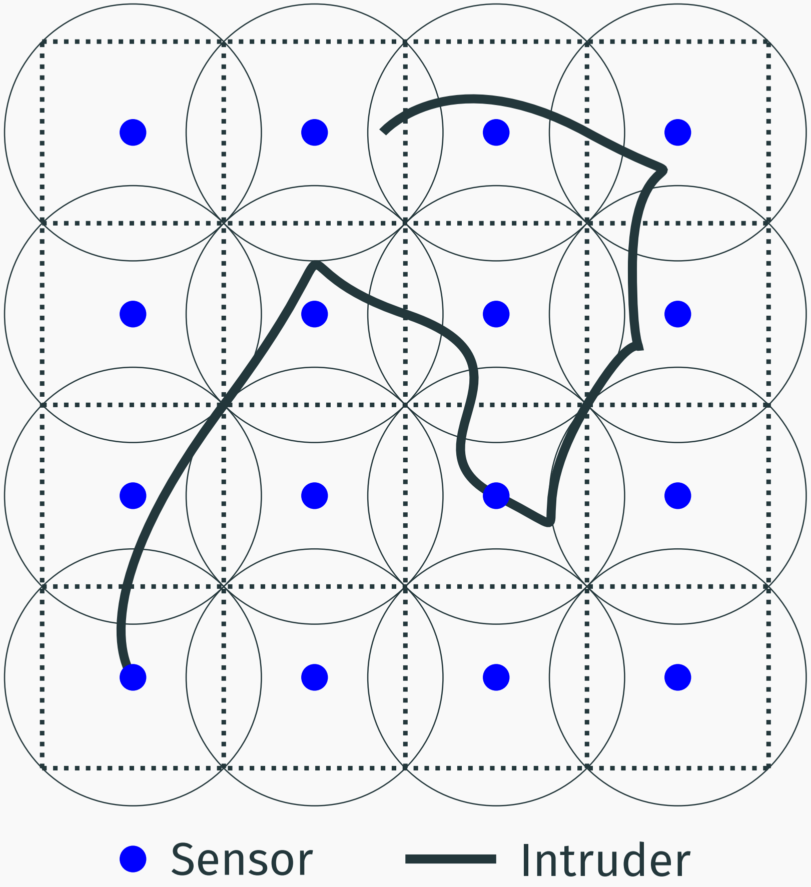
Sensor loops at two points along the road



How to switch traffic lights given L1 and L2?

How to choose L1 and L2 for a given policy and road network?

Application III: Intrusion detection using sensor networks



Aim:

- minimize the energy consumption of the sensors, while
- keeping tracking error to a minimum

Common application traits

Stochastic:
noisy observations

Model-free:
sample access to objective
* gradients unavailable

High-dimensional:
brute-force search infeasible

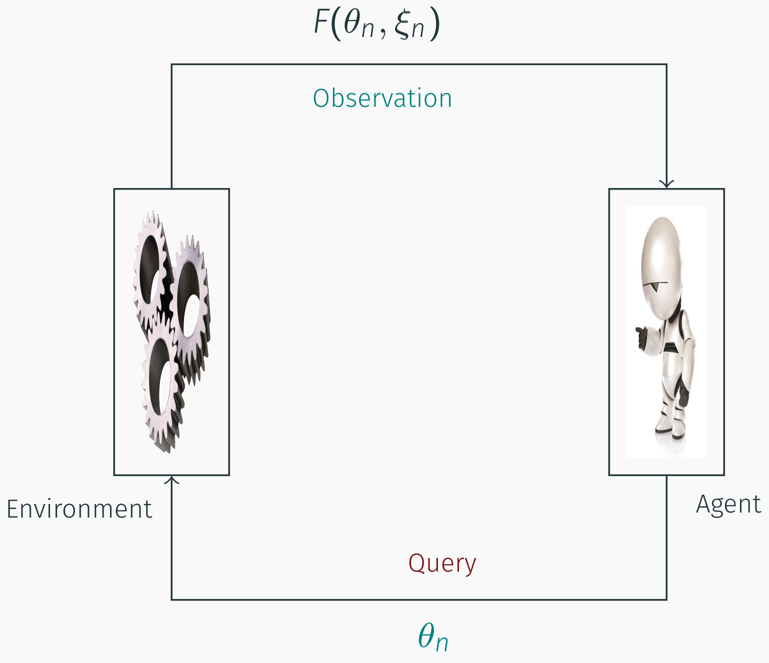
Solution:
Simultaneous perturbation methods

The framework

Basic optimization problem

noisy observation

$F(\theta_n, \xi_n)$



θ_n

Aim: $\theta^* = \underset{\theta \in \Theta}{\arg\min} \left\{ f(\theta) \triangleq \mathbb{E}[F(\theta, \xi)] \right\},$

- f: $\mathbb{R}^{\overset{d}{N}} \to \mathbb{R}$ is the performance measure
    - f *not* assumed to be convex
- $F(\theta, \xi)$ is the sample performance
- ξ is the noise factor that captures stochastic nature of the problem
- θ is the (vector) parameter of interest
- $\Theta \subseteq \mathbb{R}^{\overset{d}{N}}$ is the feasible region in which θ takes values.

Stochastic optimization via simulation

Stochastic optimization deals with highly nonlinear and high dimensional systems. The challenges with these problems are:

- Too complex to solve analytically.
- Many simplifying assumptions are required.

A good alternative of modeling and analysis is ”Simulation”

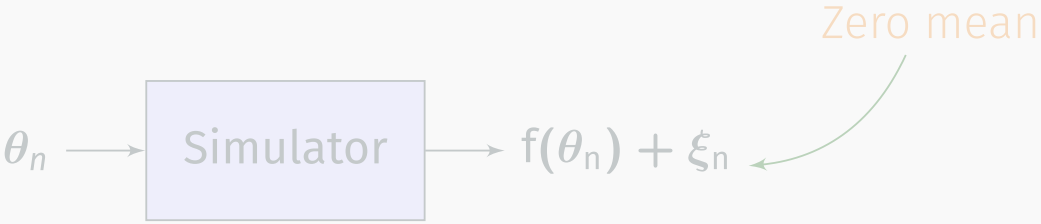


Figure 1: Simulation optimization

Stochastic optimization via simulation

Stochastic optimization deals with highly nonlinear and high dimensional systems. The challenges with these problems are:

- Too complex to solve analytically.
- Many simplifying assumptions are required.

A good alternative of modeling and analysis is ”Simulation”

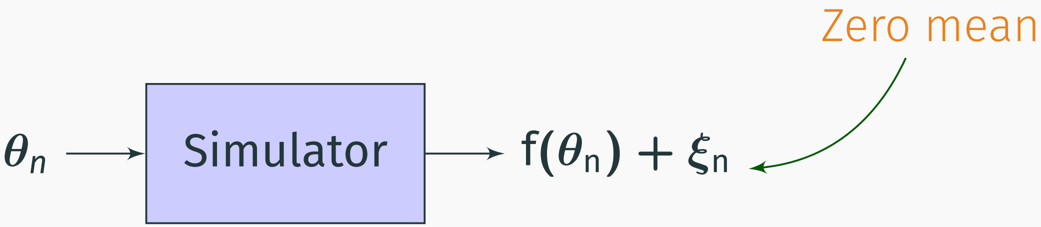


Figure 1: Simulation optimization

Noise controls

Recall: $f(\theta) = \mathbb{E}\left[F(\theta, \xi)\right]$.

Two settings for noise:

Controlled noise	ξ can be kept fixed between queries to obtain $F(\theta_1, \xi)$ and $F(\theta_2, \xi)$
Uncontrolled noise	$F(\theta, \xi)$ can be obtained at any point, but ξ is not controllable

Noise controls

Recall: $f(\theta) = \mathbb{E}\left[F(\theta, \xi)\right]$.

Two settings for noise:

Controlled noise ξ can be kept fixed between queries to obtain $F(\theta_1, \xi)$ and $F(\theta_2, \xi)$

Uncontrolled noise $F(\theta, \xi)$ can be obtained at any point, but ξ is not controllable

Challenges in simulation optimization

Deterministic optimization problem

- focus is on search for better solutions
- Complete information about objective function f, esp. gradients

Stochastic optimization problem

- f cannot be obtained directly, but we are given sample access, i.e.,

  $$f(\theta) \equiv E_\xi[F(\theta, \xi)]$$

- Each sample $F(\theta, \xi)$ is obtained from an expensive simulation experiment or a (real) field test
- focus is on both search and evaluation
  - Tradeoff between evaluating better vs. finding more candidate solutions

Challenge: to find $\theta^* = \arg\min_{\theta \in \Theta} f(\theta)$, given only noisy function evaluations.

Challenges in simulation optimization

Deterministic optimization problem

- focus is on search for better solutions
- Complete information about objective function f, esp. gradients

Stochastic optimization problem

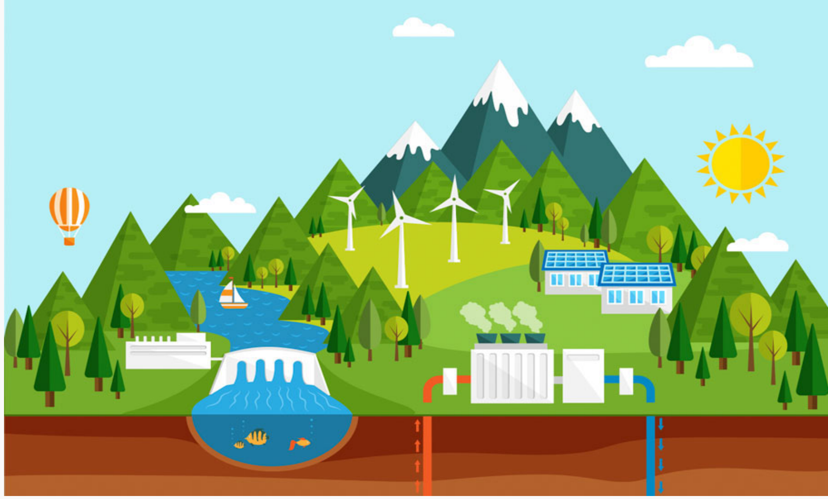
- f cannot be obtained directly, but we are given sample access, i.e., $f(\theta) \equiv E_{\xi}[F(\theta, \xi)]$
- Each sample $F(\theta, \xi)$ is obtained from an expensive simulation experiment or a (real) field test
- focus is on both search and evaluation
  - Tradeoff between evaluating better vs. finding more candidate solutions

Challenge: to find $\theta^* = \arg\min_{\theta \in \Theta} f(\theta)$, given only noisy function evaluations.

Some more applications

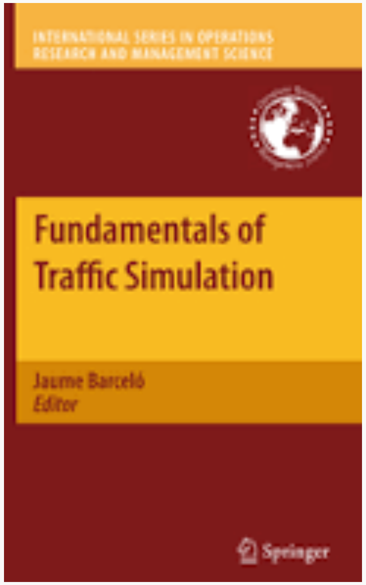
Energy Demand management

- Consumer demand, energy generation are uncertain.
- Objective is to minimize the difference.



Transportation

- Car-following model
- route choice
- traffic assignment model



INTERNATIONAL SERIES IN OPERATIONS RESEARCH AND MANAGEMENT SCIENCE

Fundamentals of Traffic Simulation

Jaume Barceló
Editor

Springer

and some more..

Service systems (banks, restaurants, call centers, amusement parks)

and some more..

Transportation systems (airports: air space, runways, baggage, roads, queues)

and here the list ends..ask INFORMS or attend WSC for more...

Manufacturing

Semiconductor fab

Supply chains

Networks

Finance

Insurance

Education

Healthcare

Banking

Mining

Oil & Gas

Call centers

Automotive OEM

Aerospace

Retirement planning

Some vendors...

- aGPSS
- Analytic solver
- Analytica
- AnyLogic
- FlexSim
- ExtendSim Pro
- Arena
- MedModel Opt Suite
- Oracle Crystal Ball
- Pedestrian dynamics
- Polaris
- ProModel Opt Suite
- SLIM
- Solver SDK Platform
- Vanguard
- Tecnomatix
- Simio
- DiscoverSim

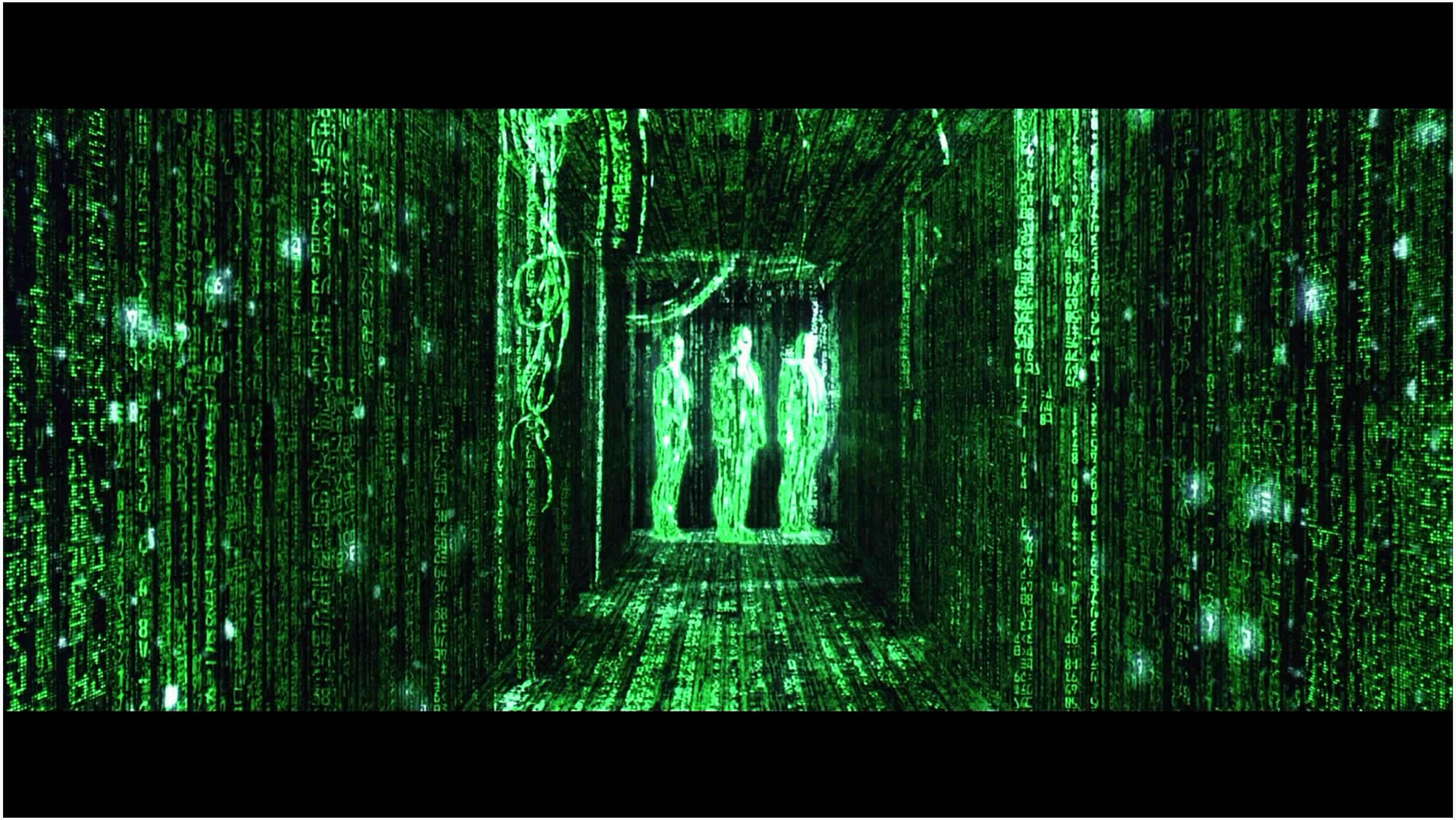
[1] James J. Swain, "Simulation Software Survey — Simulation Takes Over: Reality is for Sissies," *OR/MS Today*, Oct 2017.

Success stories...

- Kroger (Edelman 2013 finalist, gradient-based) Kroger Uses Simulation-Optimization to Improve Pharmacy Inventory Management
  - `www.youtube.com/watch?v=BNyDbBy-KYY` (start at 0:45)
  - `https://www.informs.org/About-INFORMS/News-Room/Press-Releases/Edelman-2013-Announcement`

    The Franz Edelman Award recognizes outstanding examples of innovative operations research and analytics that improves organizations and often change people's lives.
- Financial engineering
  - Monte Carlo simulation used widely on Wall Street.
  - Gradient estimates needed for hedging.
  - Hot research area: several research papers continue to be published

The Matrix has you..



Lecture 2:

First-order methods

Stochastic analog of gradient descent

$$\theta_{n+1} = \theta_n - a_n G_n. \tag{1}$$

Suppose that

- G_n is an noisy estimate of the gradient $\nabla f(\theta_n)$, i.e., $\mathbb{E}(G_n) = \nabla f(\theta_n)$.
- $\{a_n\}$ are pre-determined step-sizes satisfying:
  $$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty$$
- iterates are stable: $\sup_n \|\theta_n\| < \infty$.

Theorem (Variant of Robbins Monro stochastic approximation)
Letting $K := \{\theta \mid \nabla f(\theta) = 0\}$, we have

$$\theta_n \to K \text{ a.s. as } n \to \infty.$$

Stochastic analog of gradient descent

$$\theta_{n+1} = \theta_n - a_n G_n. \tag{1}$$

Suppose that

- G_n is an noisy estimate of the gradient $\nabla f(\theta_n)$, i.e., $\mathbb{E}(G_n) = \nabla f(\theta_n)$.
- $\{a_n\}$ are pre-determined step-sizes satisfying: $$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty$$
- iterates are stable: $\sup_n \|\theta_n\| < \infty$.

Theorem (Variant of Robbins Monro stochastic approximation)

Letting $K := \{\theta \mid \nabla f(\theta) = 0\}$, we have

$$\theta_n \to K \text{ a.s. as } n \to \infty.$$

Stochastic analog of gradient descent

$$\theta_{n+1} = \theta_n - a_n G_n. \tag{1}$$

Suppose that

- G_n is an noisy estimate of the gradient $\nabla f(\theta_n)$, i.e., $\mathbb{E}(G_n) = \nabla f(\theta_n)$.
- $\{a_n\}$ are pre-determined step-sizes satisfying:
$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty$$
- iterates are stable: $\sup_n \|\theta_n\| < \infty$.

Theorem (Variant of Robbins Monro stochastic approximation)

Letting $K := \{\theta \mid \nabla f(\theta) = 0\}$, we have

$$\theta_n \to K \text{ a.s. as } n \to \infty.$$

Stochastic analog of gradient descent

→ gradient estimate

$$\theta_{n+1} = \theta_n - a_n G_n. \tag{1}$$

→ stepsize

Suppose that

- G_n is an noisy estimate of the gradient $\nabla f(\theta_n)$, i.e., $\mathbb{E}(G_n) = \nabla f(\theta_n)$. → unbiased gradients
- $\{a_n\}$ are pre-determined step-sizes satisfying:

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty$$

e.g. $a_n = \frac{1}{n}$

- iterates are stable: $\sup_n \|\theta_n\| < \infty$.
- f is smooth

Theorem (Variant of Robbins Monro stochastic approximation)

Letting $K := \{\theta \mid \nabla f(\theta) = 0\}$, *we have*

$$\theta_n \to K \text{ a.s. as } n \to \infty.$$

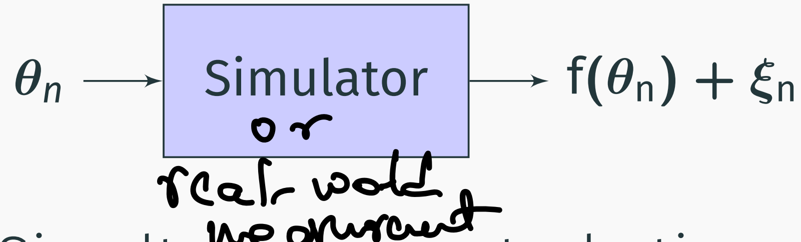
$$\theta_{n+1} = \theta_n - a_n G_n. \tag{2}$$

How to keep iterates stable?

Project θ_n onto a compact and convex set Θ ← Projected stochastic approximation

$$\theta_{n+1} = \theta_n - a_n G_n. \tag{2}$$

How to estimate the gradient of f from samples?



θ_n → Simulator → $f(\theta_n) + \xi_n$

or real-world measurement

Simultaneous perturbation methods.

Stochastic approximation (SA) alphabet soup

FDSA	Finite difference stochastic approximation
SPSA	Simultaneous perturbation stochastic approximation
SFSA	Smoothed functional stochastic approximation
RDSA	Random direction stochastic approximation

In the next few slides . . .

$$\theta_{n+1} = \theta_n - a_n G_n. \tag{3}$$

Q1) How to form G_n from function samples so that $G_n \approx \nabla f(\theta_n)$

Q2) Such a G_n - is it unbiased?

Q3) Does θ_n converge to θ^* with such a G_n?

Q4) If answer is yes to above, what is the convergence rate?

Outline

Motivation

The framework

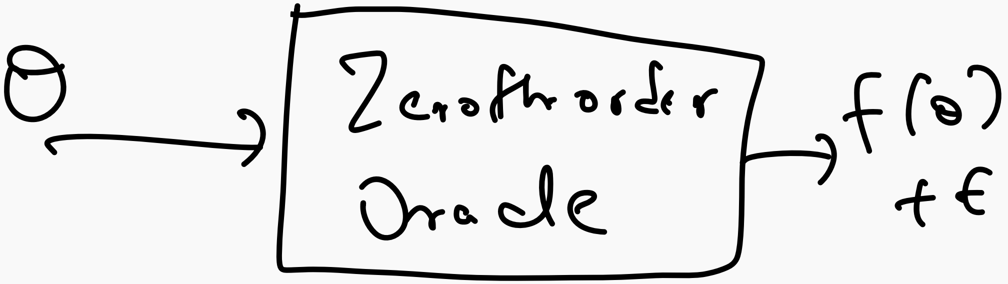
First-order methods

- How are Gradients Estimated?
- Analysis

Commercials

Second-order methods

Applications



For δ small,

$$f'(\theta) \approx \frac{f(\theta+\delta)-f(\theta)}{\delta}$$

Finite-differencing

Perfect measurements ⇔ No noise

$\theta \rightarrow \boxed{} \rightarrow f(\theta)$

Finite-difference stochastic approximation (FDSA) (Kiefer and Wolfowitz, 1952):

One-sided gradient estimate →

$$g^i = \frac{1}{\delta}\left(f(\theta + \delta e_i) - f(\theta)\right), \quad i = 1, \dots, \cancel{N}\, d$$

Assume $f \in \mathcal{C}^3$

Taylor-series expansion:

$$f(\theta + \delta e_i) = f(\theta) + \delta\, \underbrace{\nabla f(\theta)^\top e_i}_{= \nabla_i f(\theta)} + \frac{\delta^2}{2} e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

Assume Hessian norm is bounded above

Accuracy: $\|g - \nabla f(\theta)\|_2 = O(\delta)$.

Suppose $\exists\, \tilde{g}$ s.t. $\|\tilde{g} - \nabla f(\theta)\| = O(\delta^2)$

Needs $N + 1$ queries.

Perfect measurements ⇔ No noise

Finite-difference stochastic approximation (FDSA) (Kiefer and Wolfowitz, 1952):

$$g^i = \frac{1}{\delta}\left(f(\theta + \delta e_i) - f(\theta)\right), \quad i = 1, \ldots, \cancel{N} d.$$

Assume $f \in \mathcal{C}^3$ (three times continuously differentiable)

Taylor-series expansion:

$$f(\theta + \delta e_i) = f(\theta) + \delta\, \nabla f(\theta) e_i + \frac{\delta^2}{2} e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

Accuracy: $\|g - \nabla f(\theta)\|_2 = O(\delta)$.

for one gradient descent update iteration

Needs $\cancel{N} d + 1$ queries.

FDSA with two-sided Differences

Improved estimate:

Balanced estimator →

$$g^i = \frac{1}{2\delta}\left(f(\theta+\delta e_i) - f(\theta-\delta e_i)\right), \quad i = 1, \ldots, \cancel{N} d$$

Taylor-series expansions:

$$① \quad f(\theta+\delta e_i) = \cancel{f(\theta)} + \delta\, \nabla f(\theta)^{\top} e_i + \frac{\delta^2}{2} e_i^{\top} \cancel{\nabla^2 f(\theta)} e_i + O(\delta^3).$$

$$② \quad f(\theta-\delta e_i) = \cancel{f(\theta)} - \delta\, \nabla f(\theta)^{\top} e_i + \frac{\delta^2}{2} e_i^{\top} \cancel{\nabla^2 f(\theta)} e_i + O(\delta^3).$$

$$① - ② = 2\delta\, \nabla f(\theta)^{\top} e_i + O(\delta^3)$$

Accuracy: $\|g - \nabla f(\theta)\|_2 = O(\delta^2)$.

Needs 2~~N~~d queries.

FDSA with two-sided Differences

Improved estimate:

$$g^i = \frac{1}{2\delta}\left(f(\theta + \delta e_i) - f(\theta - \delta e_i)\right), \quad i = 1, \dots, \cancel{N}d.$$

Taylor-series expansions:

$$f(\theta + \delta e_i) = f(\theta) + \delta\,\nabla f(\theta) e_i + \frac{\delta^2}{2} e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

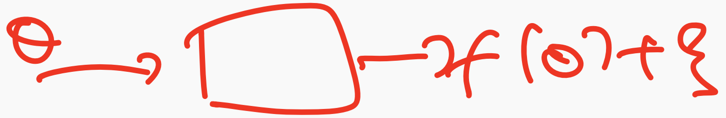
$$f(\theta - \delta e_i) = f(\theta) - \delta\,\nabla f(\theta) e_i + \frac{\delta^2}{2} e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

Accuracy: $\|g - \nabla f(\theta)\|_2 = O(\delta^2)$.

Needs 2$\cancel{N}$d queries.

FDSA + Two-sided Differences + Noise

Improved estimate:



$$G^i = \frac{1}{2\delta}\left\{f(\theta+\delta e_i)+\xi_i^+ - (f(\theta-\delta e_i)+\xi_i^-)\right\}, \quad i = 1,\ldots,N.$$

↑ zero mean noise

Taylor-series expansions:

$$f(\theta+\delta e_i) = f(\theta) + \delta\,\nabla f(\theta)^{\top} e_i + \frac{\delta^2}{2}\, e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

$$f(\theta-\delta e_i) = f(\theta) - \delta\,\nabla f(\theta)^{\top} e_i + \frac{\delta^2}{2}\, e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

Assumption: $\mathbb{E}\left[\xi^{\pm}\right] = 0$, $\mathbb{E}\left[(\xi^{\pm})\right] \le \sigma^2 < +\infty$.

$\mathbb{E}\left[G^i\right] = g^i$. Hence

$$\|\mathbb{E}[G] - \nabla f(\theta)\|_2 = O(\delta^2)\,. \longleftarrow \text{bias}$$

FDSA + Two-sided Differences + Noise

Improved estimate:

$$G^i = \frac{1}{2\delta} \left\{ f(\theta + \delta e_i) + \xi_i^+ - (f(\theta - \delta e_i) + \xi_i^-) \right\}, \quad i = 1, \ldots, N. \; d$$

Taylor-series expansions:

$$f(\theta + \delta e_i) = f(\theta) + \delta \nabla f(\theta) e_i + \frac{\delta^2}{2} e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

$$f(\theta - \delta e_i) = f(\theta) - \delta \nabla f(\theta) e_i + \frac{\delta^2}{2} e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

Assumption: $\mathbb{E}\left[\xi^\pm\right] = 0, \mathbb{E}\left[(\xi^\pm)\right] \leq \sigma^2 < +\infty.$

$\mathbb{E}\left[G^i\right] = g^i.$ Hence

$$\left\|\mathbb{E}\left[G\right] - \nabla f(\theta)\right\|_2 = O(\delta^2). \longleftarrow \textbf{bias}$$

FDSA + Two-sided Differences + Noise

Improved estimate:

$$G^i = \frac{1}{2\delta}\left\{f(\theta + \delta e_i) + \xi_i^+ - (f(\theta - \delta e_i) + \xi_i^-)\right\}, \quad i = 1, \ldots, \overset{d}{N}.$$

Taylor-series expansions:

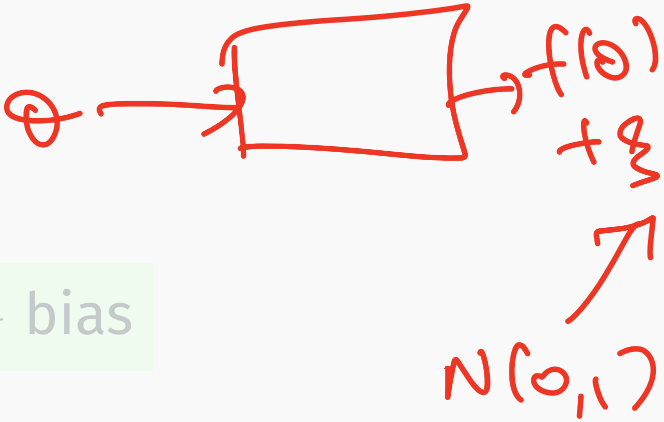
$$f(\theta + \delta e_i) = f(\theta) + \delta\, \nabla f(\theta) e_i + \frac{\delta^2}{2}\, e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

$$f(\theta - \delta e_i) = f(\theta) - \delta\, \nabla f(\theta) e_i + \frac{\delta^2}{2}\, e_i^\top \nabla^2 f(\theta) e_i + O(\delta^3).$$

Assumption: $\mathbb{E}\left[\xi^\pm\right] = 0$, $\mathbb{E}\left[(\xi^\pm)\right] \le \sigma^2 < +\infty$.

$\mathbb{E}\left[G^i\right] = g^i$. Hence

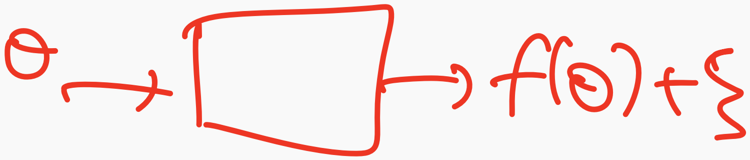
$$\|\mathbb{E}\left[G\right] - \nabla f(\theta)\|_2 = O(\delta^2)\,. \longleftarrow \text{bias}$$



FDSA + Two-sided Differences + Noise

Improved estimate:

$$G^i = \frac{1}{2\delta}\left\{f(\theta+\delta e_i)+\xi_i^{+} - \left(f(\theta-\delta e_i)+\xi_i^{-}\right)\right\}, \quad i = 1,\ldots,\cancel{N}d.$$



Taylor-series expansions:

$$f(\theta+\delta e_i) = f(\theta) + \delta\,\nabla f(\theta)e_i + \frac{\delta^2}{2}\,e_i^{\top}\nabla^2 f(\theta)e_i + O(\delta^3).$$

$$f(\theta-\delta e_i) = f(\theta) - \delta\,\nabla f(\theta)e_i + \frac{\delta^2}{2}\,e_i^{\top}\nabla^2 f(\theta)e_i + O(\delta^3).$$

Assumption: $\mathbb{E}\left[\xi^{\pm}\right] = 0$, $\mathbb{E}\left[(\xi^{\pm})^{2}\right] \le \sigma^2 < +\infty$.

$\mathbb{E}\left[G^i\right] = g^i$. Hence

$$\|\mathbb{E}[G] - \nabla f(\theta)\|_2 = O(\delta^2)\,. \longleftarrow \textbf{bias}$$

So far: with FDSA, we can get a gradient estimate

$$G^i = \frac{1}{2\delta}\left\{f(\theta + \delta e_i) + \xi_i^+ - (f(\theta - \delta e_i) + \xi_i^-)\right\}, \quad i = 1, \ldots, \cancel{N}d$$ with

bias $O(\delta^2)$

what is second moment: $\mathbb{E}\left[\|G\|_2^2\right] = ?$

$G_i = g_i + \frac{\xi_i^+ - \xi_i^-}{2\delta}$, hence $\mathbb{E}\left[G_i^2\right] = g_i^2 + \frac{2\sigma^2}{4\delta^2} = g_i^2 + \frac{\sigma^2}{2\delta^2}$ and

$$\mathbb{E}\left[\|G\|_2^2\right] = \|g\|_2^2 + O\left(\frac{N}{\delta^2}\right).$$

So far: with FDSA, we can get a gradient estimate

$$G^i = \frac{1}{2\delta}\left\{f(\theta + \delta e_i) + \xi_i^+ - (f(\theta - \delta e_i) + \xi_i^-)\right\}, \quad i = 1, \ldots, \cancel{N} d$$

with bias $O(\delta^2)$

what is second moment: $\mathbb{E}\left[\|G\|_2^2\right] = ?$

$G_i = g_i + \frac{\xi_i^+ - \xi_i^-}{2\delta}$, hence $\mathbb{E}\left[G_i^2\right] = g_i^2 + \frac{2\sigma^2}{4\delta^2} = g_i^2 + \frac{\sigma^2}{2\delta^2}$ and

$$\mathbb{E}\left[\|G\|_2^2\right] = \|g\|_2^2 + O\left(\frac{N}{\delta^2}\right).$$

So far: with FDSA, we can get a gradient estimate

$$G^i = \frac{1}{2\delta}\left\{f(\theta + \delta e_i) + \xi_i^+ - (f(\theta - \delta e_i) + \xi_i^-)\right\}, \quad i = 1, \ldots, \overset{d}{N}$$ with

bias $O(\delta^2)$

what is second moment: $\mathbb{E}\left[\|G\|_2^2\right] = ?$

$G_i = g_i + \frac{\xi_i^+ - \xi_i^-}{2\delta}$, hence $\mathbb{E}\left[G_i^2\right] = g_i^2 + \frac{2\sigma^2}{4\delta^2} = g_i^2 + \frac{\sigma^2}{2\delta^2}$ and

$$\mathbb{E}\left[\|G\|_2^2\right] = \|g\|_2^2 + O\left(\frac{\overset{d}{N}}{\delta^2}\right).$$

$$\|\mathbb{E}G - \nabla f\| = O(\delta^2)$$

So far: with FDSA, we can get a gradient estimate

$$G^i = \frac{1}{2\delta}\left\{f(\theta+\delta e_i)+\xi_i^+ - (f(\theta-\delta e_i)+\xi_i^-)\right\}, \quad i = 1,\dots,\cancel{N}\,d.$$ with

bias $O(\delta^2)$

$$G = \begin{pmatrix} G_1 \\ \vdots \\ G_d \end{pmatrix}$$

what is second moment: $\mathbb{E}\left[\|G\|_2^2\right] =?$

$$G_i = g_i + \frac{\xi_i^+ - \xi_i^-}{2\delta}, \text{ hence } \mathbb{E}\left[G_i^2\right] = g_i^2 + \frac{2\sigma^2}{4\delta^2} = g_i^2 + \frac{\sigma^2}{2\delta^2}$$ and

$$\mathbb{E}\left[\|G\|_2^2\right] = \|g\|_2^2 + O\left(\frac{\cancel{N}\,d}{\delta^2}\right). \Rightarrow \mathrm{Var}(G) = O\left(\frac{d}{\delta^2}\right)$$

because $\mathbb{E}(\xi_i^+ - \xi_i^-) = 0$

$$\mathbb{E}G_i^2 = \mathbb{E}g_i^2 + \mathbb{E}\left(\frac{(\xi_i^+ - \xi_i^-)^2}{4\delta^2}\right) + 2\frac{g_i}{2\delta}\,\cancel{\mathbb{E}(\xi_i^+ - \xi_i^-)}^{\,0}$$

FDSA perturbed dimensions one-at-a-time, leading to 2N d queries.
Can we reduce the number of queries?

Idea: Simultaneously randomly perturb all dimensions! (Spall, 1992)

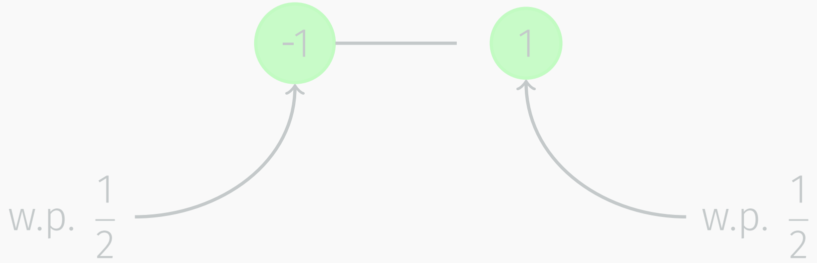
Function measurements

$$y_n^+ = f(\theta_n + \delta_n d_n) + \xi_n^+, \quad y_n^- = f(\theta_n - \delta_n d_n) + \xi_n^-$$

Gradient estimate

$$G^i = \left[\frac{y_n^+ - y_n^-}{2\delta_n d_n^i}\right].$$

How to choose $d_n^i, i = 1, \ldots, N$ d?



Only 2-queries, regardless of N!

$\mathbb{E}\left[G^i\right] = g^i$! Hence, $\|\mathbb{E}[G] - \nabla f(\theta)\|_2 = O(\delta^2)$.

FDSA perturbed dimensions one-at-a-time, leading to 2~~N~~ d queries.
Can we reduce the number of queries?

Idea: Simultaneously randomly perturb all dimensions! (Spall, 1992)

Function measurements

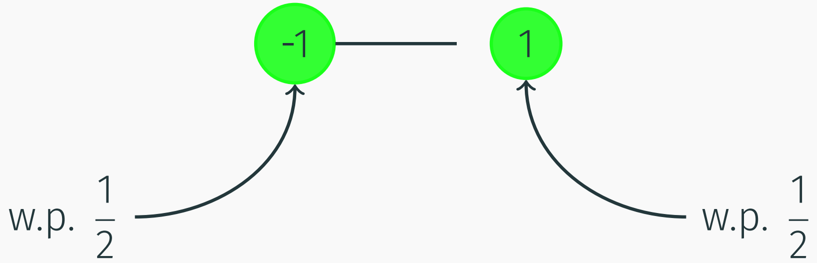
$$y_n^+ = f(\theta_n + \delta_n d_n) + \xi_n^+, \quad y_n^- = f(\theta_n - \delta_n d_n) + \xi_n^-$$

Gradient estimate

$$G^i = \left[\frac{y_n^+ - y_n^-}{2\delta_n d_n^i}\right].$$

→ Rademacher r.v.

How to choose $d_n^i, i = 1, \ldots,$ ~~N~~ d?



Only 2-queries, regardless of ~~N~~ d!

$\mathbb{E}\left[G^i\right] = g^i$! Hence, $\|\mathbb{E}[G] - \nabla f(\theta)\|_2 = O(\delta^2)$.

FDSA perturbed dimensions one-at-a-time, leading to 2~~N~~d queries.
Can we reduce the number of queries?

Idea: Simultaneously randomly perturb all dimensions! (Spall, 1992)

Function measurements

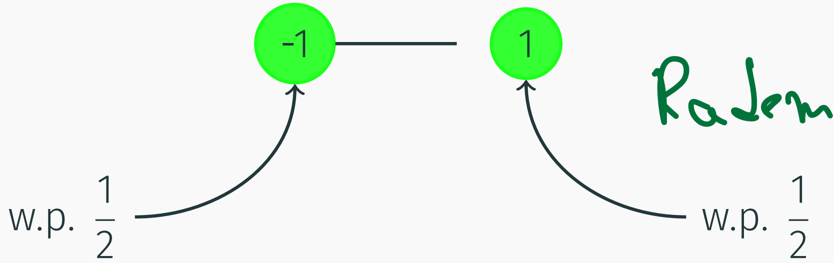
→ d_n → random vector

$$y_n^+ = f(\theta_n + \delta_n d_n) + \xi_n^+, \quad y_n^- = f(\theta_n - \delta_n d_n) + \xi_n^-$$

Gradient estimate

$$G_n^i = \left[\frac{y_n^+ - y_n^-}{2\delta_n d_n^i}\right].$$

How to choose $d_n^i, i = 1, \ldots, \cancel{N} d$?



Rademacher r.v.

$$\theta_{n+1} = \theta_n - a_n G_n$$

Only 2-queries, regardless of ~~N~~d!

$\mathbb{E}\left[G^i\right] = g^i$! Hence, $\|\mathbb{E}[G] - \nabla f(\theta)\|_2 = O(\delta^2)$.

FDSA perturbed dimensions one-at-a-time, leading to $2N$ queries.
Can we reduce the number of queries?

Idea: Simultaneously randomly perturb all dimensions! (Spall, 1992)

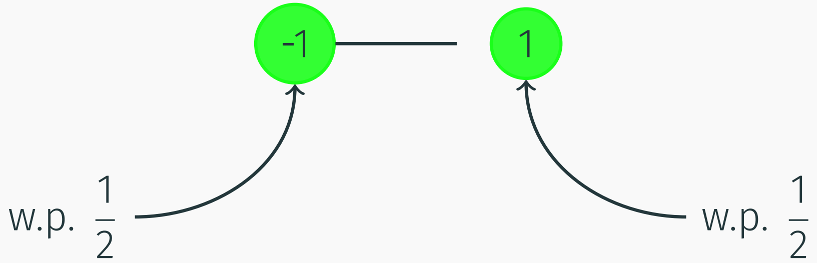
Function measurements

$$y_n^+ = f(\theta_n + \delta_n d_n) + \xi_n^+, \quad y_n^- = f(\theta_n - \delta_n d_n) + \xi_n^-$$

Gradient estimate

$$G^i = \left[\frac{y_n^+ - y_n^-}{2\delta_n d_n^i}\right].$$

How to choose $d_n^i, i = 1, \ldots, N$ d?



Only 2-queries, regardless of N d!

$\mathbb{E}\left[G^i\right] = g^i$! Hence, $\|\mathbb{E}[G] - \nabla f(\theta)\|_2 = O(\delta^2)$.

FDSA perturbed dimensions one-at-a-time, leading to 2~~N~~d queries.
Can we reduce the number of queries?

Idea: Simultaneously randomly perturb all dimensions! (Spall, 1992)

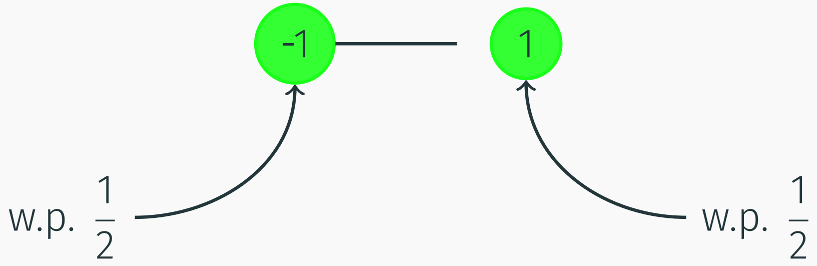
Function measurements

$$y_n^+ = f(\theta_n + \delta_n d_n) + \xi_n^+, \quad y_n^- = f(\theta_n - \delta_n d_n) + \xi_n^-$$

Gradient estimate

$$G^i = \left[\frac{y_n^+ - y_n^-}{2\delta_n d_n^i}\right].$$

How to choose $d_n^i, i = 1, \ldots, N$?



Only 2-queries, regardless of ~~N~~d!

$\mathbb{E}\left[G^i\right] = g^i$! Hence, $\|\mathbb{E}[G] - \nabla f(\theta)\|_2 = O(\delta^2)$.

Function measurements

$$y_n^+ = f(\theta_n + \delta_n d_n) + \xi_n^+, \quad y_n^- = f(\theta_n - \delta_n d_n) + \xi_n^-.$$

Gradient estimate

$$G^i = \left[\frac{y_n^+ - y_n^-}{2\delta_n d_n^i}\right].$$

Taylor series expansions

$$f(\theta_n \pm \delta_n d_n) = f(\theta_n) \pm \delta_n d_n^\top \nabla f(\theta_n) + \frac{\delta_n^2}{2} d_n^\top \nabla^2 f(\theta_n) d_n + O(\delta_n^3)$$

$$\frac{f(\theta_n + \delta_n d_n) - f(\theta_n - \delta_n d_n)}{2\delta_n d_n^i} = \nabla_i f(\theta_n) + \sum_{j=1, j\neq i}^{N} \frac{d_n^j}{d_n^i} \nabla_j f(\theta_n) + O(\delta_n^2)$$

zero-mean since d_n symmetric Bernoulli ±1 r.v.s

Hence, $\left\|\mathbb{E}\left[G^i\right] - \nabla f(\theta_n)\right\|_2 = O(\delta_n^2)$.

Function measurements

$$y_n^+ = f(\theta_n + \delta_n d_n) + \xi_n^+, \quad y_n^- = f(\theta_n - \delta_n d_n) + \xi_n^-.$$

Gradient estimate

$$G^i = \left[\frac{y_n^+ - y_n^-}{2\delta_n d_n^i}\right].$$

$$E\, G^i = E\left[\frac{f(\theta_n + \delta_n d_n) - f(\theta_n - \delta_n d_n)}{2\delta_n d_n^i}\right]$$

Taylor series expansions

$$f(\theta_n \pm \delta_n d_n) = f(\theta_n) \pm \delta_n d_n^\top \nabla f(\theta_n) + \frac{\delta_n^2}{2} d_n^\top \nabla^2 f(\theta_n) d_n + O(\delta_n^3)$$

$$\frac{f(\theta_n + \delta_n d_n) - f(\theta_n - \delta_n d_n)}{2\delta_n d_n^i} = \nabla_i f(\theta_n) + \sum_{j=1, j\neq i}^{N} \frac{d_n^j}{d_n^i} \nabla_j f(\theta_n) + O(\delta_n^2)$$

zero-mean since d_n symmetric Bernoulli ±1 r.v.s

Hence, $\left\| \mathbb{E}\left[G^i\right] - \nabla f(\theta_n) \right\|_2 = O(\delta_n^2).$

$$f(\theta_n \pm \delta_n d_n) = f(\theta_n) \pm \delta_n d_n^\top \nabla f(\theta_n) + \frac{\delta_n^2}{2} d_n^\top \nabla^2 f(\theta_n) d_n + O(\delta_n^3)$$

$$f(\theta_n+\delta_n d_n) - f(\theta_n-\delta_n d_n) = 2\delta_n d_n^\top \nabla f(\theta_n) + O(\delta_n^3)$$

$$\frac{f(\theta_n+\delta_n d_n) - f(\theta_n-\delta_n d_n)}{2\delta_n d_n^i}$$

$$= \frac{\sum_{j=1}^{d} d_n^j \nabla_j f(\theta_n)}{d_n^i} + O(\delta_n^2)$$

$$= \nabla_i f(\theta_n) + \sum_{j\neq i} \frac{d_n^j}{d_n^i} \nabla_j f(\theta_n) + O(\delta_n^2)$$

treat θ_n as a constant in this expectation

$$E\left(\frac{f(\theta_n+\delta_n d_n) - f(\theta_n-\delta_n d_n)}{2\delta_n d_n^i}\right)$$

$$= \nabla_i f(\theta_n) + \sum_{j\neq i} E\left(\frac{d_n^j}{d_n^i}\right) \nabla_j f(\theta_n) + O(\delta_n^2)$$

zero in expectation

Note!

$$d_n = (\underbrace{d_n^1 \cdots d_n^d}_{\text{independent Rademacher}})$$

$$= \nabla_i f(\theta_n) + O(\delta_n^2)$$

Function measurements

$$y_n^+ = f(\theta_n + \delta_n d_n) + \xi_n^+, \quad y_n^- = f(\theta_n - \delta_n d_n) + \xi_n^-.$$

Gradient estimate

$$G^i = \left[\frac{y_n^+ - y_n^-}{2\delta_n d_n^i}\right].$$

Taylor series expansions

$$f(\theta_n \pm \delta_n d_n) = f(\theta_n) \pm \delta_n d_n^\top \nabla f(\theta_n) + \frac{\delta_n^2}{2} d_n^\top \nabla^2 f(\theta_n) d_n + O(\delta_n^3)$$

$$\frac{f(\theta_n + \delta_n d_n) - f(\theta_n - \delta_n d_n)}{2\delta_n d_n^i} = \nabla_i f(\theta_n) + \sum_{j=1, j\neq i}^{N} \frac{d_n^j}{d_n^i} \nabla_j f(\theta_n) + O(\delta_n^2)$$

zero-mean since d_n symmetric Bernoulli ±1 r.v.s

Hence, $\left\|\mathbb{E}\left[G\right] - \nabla f(\theta_n)\right\|_2 = O(\delta_n^2)$.

Lecture-3

Consider the following estimator:

$$G = \Delta\left[\frac{\left(f(\theta+\delta\Delta)+\xi^{+}\right) - \left(f(\theta-\delta\Delta)+\xi^{-}\right)}{2\delta}\right]$$

$\Delta = (\Delta_1, \dots, \Delta_d)^T$ independent

$$\Delta_i \sim \mathcal{N}(0,1)$$

Is G a good enough estimator of $\nabla f(\theta)$?

Yes. Gaussian smoothed functional

aka Gaussian Smoothing

“Mother” of all two-point sim-pert estimates

$$G = \frac{(f(\theta + U) + \xi^{+}) - (f(\theta - U) + \xi^{-})}{2\delta} V\,.$$

Choose U, V such that $\mathbb{E}\left[VU^{\top}\right] = I$, $\mathbb{E}\left[V\right] = 0$.

One-point estimate!

$$G = \frac{(f(\theta + U) + \xi^{+})}{\delta} V\,.$$

Choose U, V such that $\mathbb{E}\left[VU^{\top}\right] = I$, $\mathbb{E}\left[V\right] = 0$. Works??

$$\mathbb{E}\left[G\right] = \mathbb{E}\left[G - \frac{f(\theta)}{\delta} V\right] = \mathbb{E}\left[\frac{(f(\theta + U) + \xi^{+}) - f(\theta)}{\delta} V\right].$$

“Mother” of all two-point sim-pert estimates

$$G = \frac{(f(\theta + U) + \xi^+) - (f(\theta - U) + \xi^-)}{2\delta} V\,.$$

Choose U, V such that $\mathbb{E}\left[VU^\top\right] = I$, $\mathbb{E}\left[V\right] = 0$.

One-point estimate!

$$G = \frac{(f(\theta + U) + \xi^+)}{\delta} V\,.$$

Choose U, V such that $\mathbb{E}\left[VU^\top\right] = I$, $\mathbb{E}\left[V\right] = 0$. Works??

$$\mathbb{E}\left[G\right] = \mathbb{E}\left[G - \frac{f(\theta)}{\delta} V\right] = \mathbb{E}\left[\frac{(f(\theta + U) + \xi^+) - f(\theta)}{\delta} V\right].$$

Family of “Sim-pert” Gradient Estimates

- $U \sim \delta\,\mathcal{N}(0, I)$, $V = \delta^{-1} U$
  - Smoothed functional by Katkovnik and Kulchitsky (1972);
  - Refined by Polyak and Tsybakov (1990); also studied by Dippon (2003); Nesterov and Spokoiny (2011).
- $U \sim \delta\,\mathrm{Unif}(\mathbb{S}_N)$, $V = N\delta^{-1} U$
  - RDSA by Kushner and Clark (1978); Enhanced by Prashanth et al. (2017)
  - Rediscovered by Flaxman et al. (2005)
- $U_i \sim \delta\,\mathrm{Rademacher}(\pm 1)$, $V = \delta^{-1} U$
  - SPSA by Spall (1992).
- Deterministic perturbations by Bhatnagar et al. (2003)
- ...

Does it matter which of these we select? Not really:
Bias is always $O(\delta^2)$, while variance is $O(1)$ or $O(\delta^{-2})$ (noise controlled or not)

Family of “Sim-pert” Gradient Estimates

- $U \sim \delta\,\mathcal{N}(0, I)$, $V = \delta^{-1} U$
  - Smoothed functional by Katkovnik and Kulchitsky (1972);
  - Refined by Polyak and Tsybakov (1990); also studied by Dippon (2003); Nesterov and Spokoiny (2011).
- $U \sim \delta\,\mathrm{Unif}(\mathbb{S}_N)$, $V = N\delta^{-1} U$
  - RDSA by Kushner and Clark (1978); Enhanced by Prashanth et al. (2017)
  - Rediscovered by Flaxman et al. (2005)
- $U_i \sim \delta\,\mathrm{Rademacher}(\pm 1)$, $V = \delta^{-1} U$
  - SPSA by Spall (1992).
- Deterministic perturbations by Bhatnagar et al. (2003)
- . . .

Does it matter which of these we select? Not really:
Bias is always $O(\delta^2)$, while variance is $O(1)$ or $O(\delta^{-2})$ (noise controlled or not)

Family of "Sim-pert" Gradient Estimates

- $U \sim \delta\,\mathcal{N}(0, I)$, $V = \delta^{-1}\,U$
  - Smoothed functional by Katkovnik and Kulchitsky (1972);
  - Refined by Polyak and Tsybakov (1990); also studied by Dippon (2003); Nesterov and Spokoiny (2011).
- $U \sim \delta\,\mathrm{Unif}(\mathbb{S}_N)$, $V = N\delta^{-1}\,U$
  - RDSA by Kushner and Clark (1978); Enhanced by Prashanth et al. (2017)
  - Rediscovered by Flaxman et al. (2005)
- $U_i \sim \delta\,\mathrm{Rademacher}(\pm 1)$, $V = \delta^{-1}\,U$
  - SPSA by Spall (1992).
- Deterministic perturbations by Bhatnagar et al. (2003)
- $\ldots$

Does it matter which of these we select? Not really:
Bias is always $O(\delta^2)$, while variance is $O(1)$ or $O(\delta^{-2})$ (noise controlled or not)

$y^+ = f(\theta+\delta U)+\xi^+$, and $y^- = f(\theta-\delta U)+\xi^-$

$$G = \frac{(y^+ - y^-)V}{2\delta}$$

Assumptions:

A2.1. Let U, V be random N-vectors satisfying $\mathbb{E}\left[VU^\top\right] = I$, $\mathbb{E}\left[V\right] = 0$, and $\mathbb{E}\left[\|V\|\,\|U\|^3\right] < \infty$.

A2.2. The noise factors $\xi^\pm$ in (2.6) satisfy

$$\mathbb{E}[\xi^+ - \xi^- \,|\, U, V] = 0, \quad \text{and} \quad \mathbb{E}[(\xi^+ - \xi^-)^2 \,|\, U, V] \leq \sigma^2 < \infty \,. \tag{2.7}$$

A2.3. The objective f satisfies

$$\sup_{\theta\in\mathbb{R}^d} \mathbb{E}[f(\theta \pm \delta U)^2] \leq B < \infty \,. \tag{2.8}$$

A2.4 f is three-times continuously differentiable with $\left|\nabla^3_{i_1 i_2 i_3} f(\theta)\right| < \bar{B} < \infty$ $\forall i_1, i_2, i_3$.

Claim: Assume A2.1 – 2.4. Then,

$$\| \mathbb{E}\, G - \nabla f(\theta) \| \leq C_1 \delta^2 \quad \text{and}$$

$$\mathbb{E}\left[\| G - \mathbb{E} G \|^2\right] \leq \frac{C_2}{\delta^2}$$

Proof:

$$E G = E\left[v \left(\frac{f(\theta+\delta u) - f(\theta-\delta u)}{2\delta} \right)\right]$$

(Since $E\left(v \left(\frac{\xi^{+} - \xi^{-}}{2\delta} \right)\right) = \frac{1}{2\delta} E\left[E\left[v(\xi^{+} - \xi^{-}) \mid v \right]\right] = E\left[v \, E\left[\frac{\xi^{+} - \xi}{2\delta} \mid v \right]\right]$

$= 0$

Using Taylor series expansions,

$$f(\theta \pm \delta u) = f(\theta) \pm \delta u^T \nabla f(\theta) + \frac{\delta^2}{2} u^T \nabla^2 f(\theta) u + O(\delta^3)$$

$$v \left(\frac{f(\theta+\delta u) - f(\theta-\delta u)}{2\delta} \right) = v u^T \nabla f(\theta) + O(\delta^2)$$

Taking expectations,

using $E(v u^T) = I$, we get

$$\| E G - \nabla f(\theta) \| \le c_1 \delta^2$$

Second claim proof:

$$E\|G - EG\|^2 \le 4E\|G\|^2$$

$$= 4E\left[\|V\|^2\left\{\left(\frac{\xi^+-\xi^-}{2\delta}\right)^2 + 2\overbrace{\left(\frac{\xi^+-\xi^-}{2\delta}\right)\left(\frac{f(\theta+\delta U)-f(\theta-\delta U)}{2\delta}\right)}^{\text{near zero}} + \left(\frac{f(\theta+\delta U)-f(\theta-\delta U)}{2\delta}\right)^2\right\}\right]$$

$$= 4E\left[\frac{\|V\|^2(\xi^+-\xi^-)^2}{4\delta^2}\right] + 4E\left[\frac{\|V\|^2(f(\theta+\delta U)-f(\theta-\delta U))^2}{4\delta^2}\right]^2$$

$$\le C_2/\delta^2$$

$\downarrow$

Using (i) $E\, f(\theta \pm \delta U)^2 \le B_1 < \infty$

(ii) $E\|V\|^2 \le B_2 < \infty$

What we have learned so far?

For performing gradient descent:

$$\theta_{n+1} = \theta_n - a_n G_n,$$

we can construct nearly unbiased gradient estimate G_n using simultaneous perturbation trick

Noise → Gradient estimate ↓	Controlled	Uncontrolled
Bias	“$C_1 \delta$” $C_1\delta$	$C_1\delta^2$
Variance	C_2	$\frac{C_2}{\delta^2}$

to be handled later.

This assumed $f \in \mathcal{C}^3$. Holds also for f convex, smooth.

A few answers so far. . .

Q1) How to form G_n from function samples so that $G_n \approx \nabla f(\theta_n)$

Use simultaneous perturbation trick

Q2) Such a G_n - is it unbiased?

Almost . . . what we get is an asymptotically unbiased estimate?

Q3) Does $\theta_{n+1} = \theta_n - a_n G_n$ converge to θ^* with such a G_n?

??

Q4) If answer is yes to above, what is the convergence rate?

??

A few answers so far. . .

Q1) How to form G_n from function samples so that $G_n \approx \nabla f(\theta_n)$

Use simultaneous perturbation trick

Q2) Such a G_n - is it unbiased?

Almost . . . what we get is an asymptotically unbiased estimate? if we let $\delta_n \to 0$ since bias is $O(\delta_n^2)$

Q3) Does $\theta_{n+1} = \theta_n - a_n G_n$ converge to θ^* with such a G_n?

??

Q4) If answer is yes to above, what is the convergence rate?

??

A few answers so far...

Q1) How to form G_n from function samples so that $G_n \approx \nabla f(\theta_n)$

Use simultaneous perturbation trick

Q2) Such a G_n - is it unbiased?

Almost .... what we get is an asymptotically unbiased estimate?

→ Stationary point $\nabla f(\theta^*) = 0$

Q3) Does $\theta_{n+1} = \theta_n - a_n G_n$ converge to θ^* with such a G_n?

??

Q4) If answer is yes to above, what is the convergence rate?

??

Can I bound

$E\|\theta_n - \theta^*\|^2$ or

$E\left(f(\theta_n) - f(\theta^*)\right)$ or $E\|\nabla f(\theta_n)\|^2$

Outline

Motivation

The framework

First-order methods

How are Gradients Estimated?

Analysis → To be covered at a later point in the course after introducing the necessary background on stochastic approximation

Commercials

Second-order methods

Applications

xkcd on real analysis

THERE EXISTS SOME NUMBER X SUCH THAT f(x)=G(f(0))=1.
OH YES.
SOMEWHERE OUT THERE, IT EXISTS.
AND WE MUST FIND IT... AND DESTROY IT.
GRAB YOUR SWORDS, STUDENTS! WE RIDE!
I THINK I'M IN THE WRONG MATH CLASS?
I'M FINALLY IN THE RIGHT ONE.