

# Stochastic Gradient Techniques for Optimization and Learning

Felisa J. Vázquez-Abad and Bernd Heidergott

University of Melbourne 2019

[Version: March 6, 2019]

---

# Contents

<b>I</b>	<b>Theory of Stochastic Optimization and Learning</b>	<b>9</b>
<b>1</b>	<b>Deterministic Optimization</b>	<b>11</b>
1.1	Unconstrained Optimization . . . . .	11
1.2	Constrained Optimization . . . . .	24
1.3	Practical Considerations . . . . .	34
1.4	Exercises . . . . .	36
<b>2</b>	<b>The Iterative Method as an ODE</b>	<b>39</b>
2.1	Motivation . . . . .	39
2.2	Stability of ODE's . . . . .	41
2.3	ODE limit of recursive algorithms . . . . .	46
2.4	ODE method for Optimization and Learning . . . . .	53
2.5	Exercises . . . . .	57
<b>3</b>	<b>Stochastic Approximation, Exogenous Noise Model</b>	<b>59</b>
3.1	Motivation . . . . .	59
3.2	The Robbins Monro Procedure . . . . .	61
3.3	Exogenous noise model, decreasing stepsize . . . . .	63
3.4	Summary for the Exogenous Noise Case . . . . .	72
3.5	Exercises . . . . .	73
<b>4</b>	<b>Stochastic Approximation, Endogenous Noise Model</b>	<b>77</b>
4.1	The Endogenous Noise Model . . . . .	77
4.2	Constant Step Size, Weak Convergence . . . . .	81
4.3	Exercises . . . . .	88
<b>5</b>	<b>Asymptotic Efficiency</b>	<b>91</b>
5.1	Motivation . . . . .	91
5.2	Functional CLT . . . . .	91
5.3	Estimating Confidence Intervals . . . . .	101
5.4	Asymptotic Efficiency . . . . .	102
5.5	Exercises . . . . .	104

<b>II</b>	<b>Gradient Estimation</b>	<b>107</b>
<b>6</b>	<b>A Primer for Gradient Estimation</b>	<b>109</b>
6.1	Motivation . . . . .	109
6.2	One Dimensional Distributions . . . . .	110
6.2.1	Infinitesimal Perturbation Analysis . . . . .	110
6.2.2	Score Function . . . . .	114
6.2.3	Measured Valued Differentiation . . . . .	116
6.3	A Taxonomy of Gradient Estimation . . . . .	118
6.3.1	The Static Problem . . . . .	119
6.3.2	The Random Horizon Problem . . . . .	121
6.3.3	The Steady-State Problem . . . . .	122
6.3.4	Markov Processes: The Stationary Problem . . . . .	123
6.4	Exercises . . . . .	124
<b>7</b>	<b>Gradient Estimation for the Static Problem</b>	<b>127</b>
7.1	Perturbation Analysis: IPA and SPA . . . . .	127
7.1.1	Basic Results and Techniques . . . . .	127
7.1.2	Smoothed Perturbation Analysis . . . . .	136
★7.1.3	An Indirect Approach To Establishing Unbiasedness . . . . .	138
7.2	The Score-Function Method (SF) . . . . .	139
7.2.1	Basic Results and Techniques . . . . .	139
7.2.2	Products of Measures . . . . .	145
7.3	Measure-Valued Differentiation (MVD) . . . . .	148
7.3.1	Differentiability of Products of Measures . . . . .	153
7.3.2	Differentiability of Markov Chains . . . . .	156
7.3.3	* The Weak Differentiation Approach . . . . .	164
7.4	Exercises . . . . .	165
<b>8</b>	<b>Advanced Gradient Estimation</b>	<b>171</b>
8.1	IPA Sample Path Analysis . . . . .	171
8.1.1	The Steady-State Problem . . . . .	171
8.1.2	The Randomized Problem and the Stationary Problem . . . . .	175
8.1.3	IPA for Discrete State Space Models . . . . .	176
8.2	The Score Function for the Randomized Problem . . . . .	177
8.3	Taboo Sets . . . . .	179
8.3.1	The Operator Approach (MVD) . . . . .	181
8.4	The Stationary Problem: The Operator Approach . . . . .	185
<b>III</b>	<b>Stochastic Optimization at Work</b>	<b>189</b>
<b>9</b>	<b>An Inventory Problem</b>	<b>191</b>
<b>10</b>	<b>M/G/1 Queue Study</b>	<b>197</b>
<b>11</b>	<b>An Asset Management Problem</b>	<b>201</b>

<b>12 A Neural Network Application</b>	<b>205</b>
<b>13 The Newsvendor Problem</b>	<b>207</b>
<b>A Tools from analysis</b>	<b>209</b>
A.1 Geometric Interpretation of the Gradient . . . . .	209
A.2 A Short Intermezzo on Normed Spaces and Equicontinuity . . . . .	212
A.3 Differentiation . . . . .	213
A.4 Cesàro limits . . . . .	213
A.5 Lipschitz and Uniform Continuity . . . . .	213
A.6 Interchanging Limit and Differentiation . . . . .	214
<b>B Probability Theory</b>	<b>217</b>
B.1 Measurability and Measures . . . . .	217
B.1.1 Information Structure . . . . .	217
B.1.2 Measures . . . . .	218
B.2 Expectations and Conditioning . . . . .	221
B.3 Polish Spaces . . . . .	222
B.4 Convergence of random sequences . . . . .	222
B.4.1 Types of Convergence . . . . .	222
B.5 Weak Convergence and Norm Convergence . . . . .	226
B.6 Martingale processes . . . . .	227
B.7 Regenerative Processes . . . . .	230
<b>C Markov Chains</b>	<b>231</b>
<b>Bibliography</b>	<b>233</b>



# Preface

Throughout this monograph we assume that random variables are defined on a common underlying probability space  $(\Omega, \mathcal{F}, P)$ . Furthermore, we assume that  $\mathcal{F}$  contains all Null sets with respect to  $P$ .

By convention, we equip discrete spaces with the discrete topology and the real numbers with the usually topology. Product spaces are equipped with the product topology and, unless stated otherwise, measurable spaces are equipped with the corresponding Borel fields. If not stated otherwise, random variables are real-valued and, in line with the aforementioned conventions, measurable mappings from  $(\Omega, \mathcal{F}, P)$  onto  $(\mathbb{R}, \mathbb{B})$ , with  $\mathbb{B}$  denoting the Borel field on  $\mathbb{R}$ . Expectation of a random variable  $X$  with respect to  $\mathbb{P}$  is denoted by  $\mathbb{E}$ , i.e., we write  $\mathbb{E}[X] = \int_{\Omega} X(\omega) \mathbb{P}(d\omega)$ . To simplify notation we suppress denoting  $\omega$  when this causes no confusion. We use the symbol “ $\sim$ ” to relate random variables and their corresponding distribution, i.e., we write  $X \sim F$  if  $X$  has cumulative distribution function  $F$ . We use this notation in a similar way for measures. Equality in distribution of two random variables, say,  $X$  and  $Y$  is denoted by  $X \stackrel{\mathcal{L}}{=} Y$ .

We use the following abbreviations through:

- cdf for “cumulative distribution function”
- pdf for “probability density function”
- iid for “independently identically distributed”

