Stochastic Methods for Data Science

Designing and Analyzing Algorithms for Optimization and Sampling

Jonathan Huggins
Boston University

Draft last updated: *November 6, 2023*

Preface

This book provides an introduction to the interplay between stochastic processes and algorithms, with a focus on applications in statistics and machine learning. You will learn about core concepts and results concerning stochastic processes, then use this machinery to design and analyze algorithms. The primary applications will be to (large-scale) **stochastic optimization** and to **sampling** from complex distributions – such as Bayesian posterior distributions and energy-based models – using Markov chain Monte Carlo. In addition to statistics and machine learning, the optimization and sampling algorithms student in this book have diverse applications including to problems in computer science, physics, chemistry, ecology, biology, and operations research. As such, a strong emphasis will be placed on practical implications of the results.

In light of our application-oriented motivations, the presentation of core stochastic processes material focuses on intuitive understanding of definitions and underlying theoretical results rather than their rigorous development and proofs. This approach will allow us to begin more rapidly investigating algorithms and their properties.

Expected Background You should be comfortable with vector calculus (ideally with some exposure to ordinary differential equations), linear algebra, and undergraduate probability theory. Previous exposure to ideas from statistics or machine learning – regression, probabilistic models, Markov chain Monte Carlo, (stochastic) optimization – is very helpful but not strictly required.

Contents

Notation

Conventions

- scalars are denoted by plain lowercase letters ($a, b, c, \dots, \eta, \gamma, \dots$)
- vectors are denoted by bold lowercase letters ($\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \dots, \boldsymbol{\xi}, \boldsymbol{\zeta}, \dots$)
- matrices are denotes by bold uppercase letters ($\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \dots$)
- random variables are denoted by uppercase Roman letters ($Y, W, \dots$ or $\boldsymbol{X}, \boldsymbol{Y}, \dots$) or greek letter variants ($\vartheta, \varphi, \dots$)

Acronyms and Abbreviations

- GD: gradient descent
- SGD: stochastic gradient descent
- MCMC: Markov chain Monte Carlo
- i.i.d.: independent and identically distributed

General

- $\mathbb{1}(C)$: indicator function equal to 1 if condition C is true and equal 0 if C is false
- $x \leftarrow y$: assignment of the value y to the variable x in an algorithm
- $x := y$: x is defined to be equal to y
- $x =: y$: y is defined to be equal to x

Sets and Numbers

- $\mathbb{N}$: the set of natural numbers $\{0, 1, 2, \dots\}$
- $\mathbb{Z}$: the integers $\{\dots, -2, -1, 0, 1, 2, \dots\}$
- $\mathbb{R}$: the set of real numbers
- $\mathbb{R}_+$: the set of nonnegative real numbers $[0, \infty)$

- $\mathbb{R}^D$: the set of column vectors $\boldsymbol{u} = (u_1, \ldots, u_D)^\top$ where $u_d \in \mathbb{R}$
- $\emptyset$: the empty set
- $\mathcal{P}(\Omega)$: power set of Ω
- S^c: complement of the set S
- $[k]$: the set of k integers $\{1, \ldots, k\}$

Linear Algebra

- $^\top$: transpose
- $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$: inner product $\boldsymbol{u}^\top \boldsymbol{v}$ for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D$
- $\|\boldsymbol{u}\|_2$: Euclidean norm of a vector $\boldsymbol{u} \in \mathbb{R}^D$
- $\|\boldsymbol{A}\|_2$: spectral norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{D \times K}$

Calculus

- $\nabla \phi$ or ϕ': gradient of the real-valued function ϕ
- $\nabla^2 \phi$ or ϕ'': Hessian matrix the real-valued function ϕ
- $\phi^{(p)}$: pth derivative of the real-valued function ϕ

Probability and Measure

- $\mathbb{E}(\cdot)$: expectation
- $\mathbb{P}(\cdot)$: event probability
- $\mathcal{L}_X$: denotes the distribution (also called the law) of the random variable X
- $X \sim \mu$: the random variable X has distribution μ (i.e., $\mathcal{L}_X = \mu$)
- $\delta_{\boldsymbol{x}}$: the Dirac measure given by $\delta_{\boldsymbol{x}}(A) = \mathbb{1}(\boldsymbol{x} \in A)$
- $\xrightarrow{d}$: convergence in distribution
- $\xrightarrow{p}$: convergence in probability
- $\xrightarrow{a.s.}$: convergence almost surely (i.e., with probability 1)

Models and Loss Functions

- $\mathcal{A}$: parameter space, which will always be a subset of $\mathbb{R}^D$
- $\boldsymbol{x} \in \mathcal{A}$: parameter value
- $\mathcal{L} : \mathcal{A} \to \mathbb{R}$: loss function
- $\boldsymbol{y}_n$: the nth response or observation
- $\boldsymbol{z}_n$: the covariates (i.e., features) associated with the nth response

Optimization

- $\boldsymbol{x}_\star$: optimum
- η_k: step size at iteration k

Part I

Preliminaries

Chapter 1

Optimization and Sampling

An optimization problem consists of finding the value that minimizes or maximizes a target function. A sampling problem consists of obtaining (approximate) samples from a target distribution. A variety of running examples of both problems which will be revisited throughout the book are introduced. Gradient descent and stochastic gradient descent are introduced as general-purpose algorithms for solving optimization problems. Markov chain Monte Carlo is introduced as a general approach to solving sampling problems. Stochastic methods provide the mathematical tools for the rigorous analysis of optimization and sampling algorithms, and the design of new algorithms.

1.1 Optimization

One of the fundamental problems in data science is to fit a model, improve an algorithm, or adjust a system by maximizing or minimizing a specific objective function. In this book we will focus on examples where we have a loss function $\mathcal{L}(\boldsymbol{x})$ arising from a model parameterized by $\boldsymbol{x} \in \mathcal{A}$, where often $\mathcal{A} \subseteq \mathbb{R}^D$. If we had a reward function $\mathcal{R}(\boldsymbol{x})$ to maximize, we can equivalently minimize the loss $\mathcal{L}(\boldsymbol{x}) = -\mathcal{R}(\boldsymbol{x})$, so considering only minimization problems of the form

$$\boldsymbol{x}_\star = \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x})$$

is without any loss of generality.

First we consider some models for ***regression*** (also called ***supervised learning***), where the goal is to predict a ***response*** $y \in \mathbb{Y} \subseteq \mathbb{R}$ given a vector

of ***covariates*** (also called ***features***) $\boldsymbol{z} \in \mathbb{R}^D$. The loss being minimized depends on a labeled dataset $\mathcal{D} = \{(y_n, \boldsymbol{z}_n)\}_{n=1}^N$.

Example 1.1.1 (Linear regression)**.** *We first consider the case where $\mathbb{Y} = \mathbb{R}$. For example, the response is a child's height and the covariates are the mother's height, father's height, and child's age. A widely used approach to solving this problem is linear regression, where we define a* **regression function** *$g_{\boldsymbol{x}}(\boldsymbol{z}_n) = \boldsymbol{x}^\top \boldsymbol{z}_n$ to use for prediction. The parameter $\boldsymbol{x} \in \mathbb{R}^D$ is selected to minimize the regularized squared loss function*

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}) &= \frac{1}{N}\sum_{n=1}^N (g_{\boldsymbol{x}}(\boldsymbol{z}_n) - y_n)^2 + \lambda\|\boldsymbol{x}\|_2^2 \\ &= \frac{1}{N}\sum_{n=1}^N (\boldsymbol{x}^\top \boldsymbol{z}_n - y_n)^2 + \lambda\|\boldsymbol{x}\|_2^2,\end{aligned}$$

where $\lambda \geq 0$ controls the strength of the regularization.

Example 1.1.2 (Logistic regression)**.** *Next, we consider the common scenario where $\mathbb{Y} = \{-1, +1\}$ with $+1$ indicating a "positive" example (e.g., a patient who responds to treatment) and -1 indicating a "negative" example (e.g., a patient who doesn't respond to a treatment). The goal is to correctly classify each observation as positive or negative using the covariates. A canonical approach to such classification problems is logistic regression. The regression function is the same as linear regression but now we minimize the regularized cross-entropy loss*

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}) &= \frac{1}{N}\sum_{n=1}^N \log\{1 + e^{-y_n g_{\boldsymbol{x}}(\boldsymbol{z}_n)}\} + \lambda\|\boldsymbol{x}\|_2^2 \\ &= \frac{1}{N}\sum_{n=1}^N \log\{1 + e^{-y_n \boldsymbol{x}^\top \boldsymbol{z}_n}\} + \lambda\|\boldsymbol{x}\|_2^2.\end{aligned}$$

Hence, the goal is to choose $\boldsymbol{x}$ such that $\boldsymbol{x}^\top \boldsymbol{z}_n$ is large and positive (respectively, negative) if y_n is positive (respectively, negative).

Example 1.1.3 (Support vector machines)**.** *Support vector machines (SVMs) are very similar to logistic regression, except the hinge loss replaces the cross-*

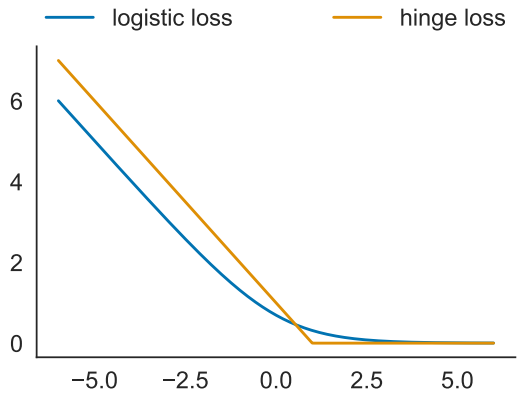


Figure 1.1: Comparison of the hinge and logistic losses. See Examples 1.1.2 and 1.1.3 for details.

entropy loss:

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}) &= \frac{1}{N}\sum_{n=1}^{N}\max(0, 1 - y_n g_{\boldsymbol{x}}(\boldsymbol{z}_n)) + \lambda\|\boldsymbol{x}\|_2^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}\max(0, 1 - y_n \boldsymbol{x}^\top \boldsymbol{z}_n) + \lambda\|\boldsymbol{x}\|_2^2.\end{aligned}$$

Figure 1.1 compares the logistic loss function $\log(1 + e^{-t})$ *to the hinge loss* $\max(0, 1 - t)$*, where* $t = y_n g_{\boldsymbol{x}}(\boldsymbol{z}_n$.

Example 1.1.4 (Artificial neural networks)**.** *Linear and logistic regression can be generalized by replacing the regression function* $g_{\boldsymbol{x}}(\boldsymbol{z}_n) = \boldsymbol{x}^\top \boldsymbol{z}_n$ *with a more flexible one. A particularly important case is an artificial neural network. Here we consider one of the simplest cases, a two-layer fully-connected network parameterized by* $\boldsymbol{x} = (\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\beta})$*, where* $\boldsymbol{W} \in \mathbb{R}^{D \times D'}$*,* $\boldsymbol{b} \in \mathbb{R}^{D'}$*, and* $\boldsymbol{\beta} \in \mathbb{R}^{D'}$*. Given a nonlinearity* $\sigma : \mathbb{R} \to \mathbb{R}$ *which we apply component-wise to a vector (e.g.,* $\phi(\boldsymbol{u}) := (\phi(u_1), \ldots, \phi(u_D))$*), the regression function is*

$$g_{\boldsymbol{x}}(\boldsymbol{z}_n) = \boldsymbol{\beta}^\top \sigma(\boldsymbol{W}^\top \boldsymbol{z}_n + \boldsymbol{b}).$$

Common choices for σ *are the rectified linear unit* $\mathrm{ReLU}(t) := \max(0, t)$*, hyperbolic tangent* $\tanh(t) = (e^{2t} - 1)/(e^{2t} + 1)$*, and the logistic sigmoid* $\varphi(t) := 1/(e^{-t} + 1)$.

Next we give some examples of ***unsupervised learning*** problems when the data consist only of observations, so dataset is $\mathcal{D} = \{\boldsymbol{y}_n\}_{n=1}^N$. The goal may be ***density estimation*** (that is, learning the distribution of the examples) or ***structure learning***, where we wish to discover interesting structure in the data. In all these cases we will first define a probabilistic model $\mathcal{M}$ consisting of distributions with densities $p_{\boldsymbol{x}}(\boldsymbol{z})$ for $\boldsymbol{x} \in \mathcal{A}$. The ***log loss*** for the dataset is then given by

$$\mathcal{L}(\boldsymbol{x}) = -\sum_{n=1}^{N} \log p_{\boldsymbol{x}}(\boldsymbol{y}_n).$$

Example 1.1.5 (Mixture models)**.** *Mixture models are used to* **cluster** *data into distinct groups. For example, a* **Gaussian mixture model** *consists of $K \geq 1$ components each with a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k \in \mathbb{R}^D$ is the mean and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{D \times D}$ is the (positive semi-definite) covariance. We will write $\mathcal{N}(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the density of a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Each component has a weight π_k representing the probability of an observation being generated by that component distribution. Thus, we enforce $w_k \geq 0$ and $\sum_{k=1}^K w_k = 1$. The model complete parameter is $\boldsymbol{x} = (\boldsymbol{w}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$ and the probability of an observation $\boldsymbol{y}$ given $\boldsymbol{x}$ is*

$$p_{\boldsymbol{x}}(\boldsymbol{y}) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{1.1}$$

Example 1.1.6 (Factor analysis)**.** *Another common type of unsupervised model aims to explain each observation as the combination of a small number of* **latent factors** $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K \in \mathbb{R}^D$*. For example, for the* **probabilistic principal component analysis** *model, let $\boldsymbol{F} \in \mathbb{R}^{D \times K}$ denote the matrix of latent factors and $\sigma > 0$ denote the noise level, so the parameter is $\boldsymbol{x} = (\boldsymbol{F}, \sigma)$. The probability of an observation $\boldsymbol{y}$ given $\boldsymbol{x}$ is*

$$p_{\boldsymbol{x}}(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{F}\boldsymbol{F}^\top + \sigma^2 \boldsymbol{I}).$$

The loss functions in both the supervised and unsupervised examples can be written in the form

$$\mathcal{L}(\boldsymbol{x}) = \sum_{n=1}^{N} \ell_{(n)}(\boldsymbol{x}) + r(\boldsymbol{x}), \tag{1.2}$$

where $\ell_{(n)}(\boldsymbol{x})$ is the loss associated with the nth observation and $r(\boldsymbol{x})$ is the regularizer. For example, for the linear regression $\ell_{(n)}(\boldsymbol{x}) = (\boldsymbol{x}^\top \boldsymbol{z}_n - y_n)^2$ and $r(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_2^2$ while for the unsupervised learning examples $\ell_{(n)}(\boldsymbol{x}) = -\log p_{\boldsymbol{x}}(\boldsymbol{z}_n)$ and $r(\boldsymbol{x}) = 0$.

Sometimes the loss can be minimized analytically.

Example 1.1.7 (Analytical solutions for linear regression and probabilistic principal component analysis)**.** *For linear regression, let $\boldsymbol{Z} \in \mathbb{R}^{N\times D}$ denote the matrix of covariate vectors and $\boldsymbol{y} = (y_1, \ldots, y_N)$ denote the vector of responses. Then $\boldsymbol{x}_\star = (\boldsymbol{Z}^\top \boldsymbol{Z} + N\lambda \boldsymbol{I})^{-1}\boldsymbol{Z}^\top \boldsymbol{y}$. Or, for probabilistic principal component analysis, let $\boldsymbol{S} = \boldsymbol{Z}^\top \boldsymbol{Z}/N$, $\boldsymbol{V} \in \mathbb{R}^{D\times K}$ the matrix whose columns are the first K eigenvectors of $\boldsymbol{S}$, and $\boldsymbol{\Lambda} \in \mathbb{R}^{K\times K}$ is the diagonal matrix of the corresponding eigenvalues, and $\lambda_{K+1}, \ldots, \lambda_D$ are the remaining eigenvalues. Then $\boldsymbol{x}_\star = (\boldsymbol{F}_\star, \sigma_\star)$, where $\sigma_\star^2 = \frac{1}{D-K}\sum_{d=K+1}^{D} \lambda_d$ and $\boldsymbol{F}_\star = \boldsymbol{V}(\boldsymbol{\Lambda} - \sigma_\star^2 \boldsymbol{I})^{1/2}$.*

However, there are two problems with relying analytic solutions. The first problem is that, in most cases, an analytic solution is not available. This is true of all remaining examples above, and many more we will encounter later. The second problem is that computing the solution analytically may be infeasible in practice, particularly when the dimension D is large. Indeed, in linear regression it requires inverting the $D \times D$ matrix $\boldsymbol{Z}^\top \boldsymbol{Z}$ while in probabilistic principal component analysis it requires computing the eigenvalues of the same matrix. Both of these procedures require $\Theta(D^3)$ operations.

So, in practice, we typically rely on *iterative numerical optimization* methods which repeatedly refine an approximation $\widehat{\boldsymbol{x}}_\star$ to the minimizer $\boldsymbol{x}_\star$ until the approximation is sufficiently accurate. In the data science context, there are four important criteria we will consider when designing optimization algorithms and evaluating their usefulness:

Optimization Algorithm Criteria

1. **General applicability.** New models and loss functions are constantly being developed. Therefore, we will focus on general-purpose optimization algorithms, rather than specialized ones that exploit distinct problem structure.

2. **Dependence on dimensionality of $\boldsymbol{x}$.** Since $\boldsymbol{x}$ is often very high-dimensional, we want methods with computational cost that is linear

or nearly linear in the dimension D. In some cases this isn't possible (e.g., if computing $\mathcal{L}(\boldsymbol{x})$ requires $\Theta(D^2)$ operations); in such cases we want the optimization algorithm to retain the same runtime dependence.

3. **Number of evaluations of $\mathcal{L}$.** When the dataset size is large or calculating $\mathcal{L}$ requires an expensive simulation (e.g., solving a system of ODEs or PDEs), it is important to evaluate $\mathcal{L}(\boldsymbol{x})$ as few times as possible.

4. **Obtaining a statistically accurate of approximation.** As seen in the examples above, the loss is a function of the observed data. Hence, we can view the loss as a noisy approximation to the *ideal* loss function we would use if given an infinite amount of data. Since the loss function is noisy, in practice it usually suffices to obtain a relatively low-precision estimate of the optimum with error of order $N^{-1/2}$.

Focusing on the case of losses in the form of Eq. (1.2), we can summarize the optimization problem we wish to solve as follows:

Problem: Finite-sum Optimization

Design general-purpose optimization algorithms for computing an estimate

$$\widehat{\boldsymbol{x}}_\star \approx \boldsymbol{x}_\star := \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{A}} \mathcal{L}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{A}} \sum_{n=1}^{N} \ell_{(n)}(\boldsymbol{x}) + r(\boldsymbol{x})$$

in a way that is scalable to high dimensions and large datasets while providing sufficient accuracy given the intrinsic noise of a finite dataset.

Since loss functions are usually differentiable, we can take advantage of the fact that the negative gradient $-\nabla\mathcal{L}(\boldsymbol{x})$ points in the direction of steepest descent. A natural algorithm for minimizing the loss using gradient information is ***gradient descent (GD)***.[1] Let $\boldsymbol{x}_0$ denote an initial parameter estimate (e.g., zero). Then for a sequence of positive step sizes $\{\eta_k\}_{k\in\mathbb{N}}$, the

[1] More sophisticated versions of gradient descent that use second-order information (e.g., Newton's method) or approximate second-order information (e.g., L-BFGS) can converge even faster. But this comes at the price of greater computational cost per iteration, often with worse dependence on the dimension.

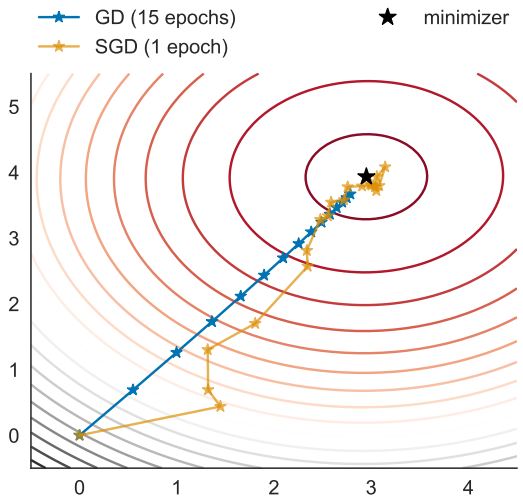


Figure 1.2: Comparison of the iterate paths of gradient descent (GD) and stochastic gradient descent (SGD) with $\eta_k = 0.1$ The contours show the loss function being optimized. See Example 1.1.8 for the details of the setup. Both algorithms are initialized at $\boldsymbol{x}_0 = (0, 0)$. Although the GD path is more "direct," SGD converges to the region near the optimum about 30 times faster.

gradient descent update is given by

$$\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k - \eta_{k+1} \nabla \mathcal{L}(\boldsymbol{x}_k). \tag{1.3}$$

Figure 1.2 shows the first few iterates of gradient descent for the problem described in Example 1.1.8 below. Gradient descent has a number of attractive features. It satisfies our first two criteria since the only requirement is that the loss is differentiable and usually the cost of computing the gradient is roughly twice that of computing the loss itself (so the runtime dependence on the dimension is essentially the best possible without assuming special structure). It converges very quickly and there are good methods for adapting the step size sequence (e.g., using a line search). However, the per-iteration computational complexity of gradient descent is $\Theta(N)$. Thus, it fails to satisfy criterion 3 and is impractical for many data science problems.

But is it really necessary to use all the data at each iteration? Or could a rough estimate of the gradient using a small amount of data be nearly as good? Recall that criterion 4 suggests we do not require an extremely accurate

approximation of $\boldsymbol{x}_\star$. So, rather than computing the gradient exactly, instead consider estimating the loss and its gradient at each iteration by restricting attention to a small "batch" of B observations selected uniformly at random from $\mathcal{D}$. At iteration k, denote the data indices that belong to the batch by $n(k,1), \ldots, n(k,B)$ and denote the batch loss as

$$\mathcal{L}_k(\boldsymbol{x}) := \frac{1}{B} \sum_{b=1}^{B} \ell_{(n(k,b))}(\boldsymbol{x}) + r(\boldsymbol{x}),$$

so $\mathbb{E}\{\mathcal{L}_k(\boldsymbol{x})\} = \mathcal{L}(\boldsymbol{x})$. Using $\mathcal{L}_k(\boldsymbol{x})$ in place of $\mathcal{L}(\boldsymbol{x})$ in the gradient descent update from Eq. (1.3) gives rise to the ***stochastic gradient descent (SGD)*** algorithm with update

$$\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k - \eta_k \nabla \mathcal{L}_k(\boldsymbol{x}_k). \tag{1.4}$$

The per-iteration iteration computational complexity of SGD is just $\Theta(B)$. Figure 1.2 illustrates the potential benefit of taking a faster but more meandering path to the minimum.

Since SGD is noisy, we might consider averaging over iterates to decrease this noise: at iteration k, define the ***iterate average*** over the most recent 50% of iterations by

$$\bar{\boldsymbol{x}}_k := \frac{1}{\lceil k/2 + 1 \rceil} \sum_{\ell=\lfloor k/2 \rfloor}^{k} \boldsymbol{x}_\ell. \tag{1.5}$$

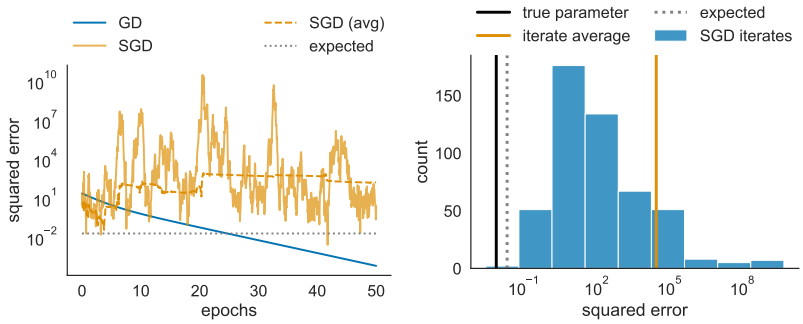
We discard the first half of the iterates since we expect early iterates to be far from the minimizer.

Example 1.1.8 (Empirical comparison of gradient descent and SGD)**.** *A simple example suffices to provide an idea for how gradient descent compares to its stochastic counterpart. Take $D = 2$ and generate $N = 100$ observations independently according to*[2]

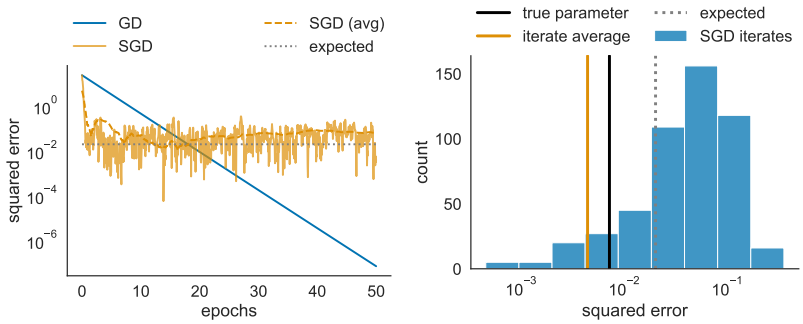
$$\boldsymbol{z}_n \sim \mathcal{N}(0, \boldsymbol{I}), \qquad y_n \mid \boldsymbol{z}_n \sim \mathcal{N}(\boldsymbol{x}_\circ^\top \boldsymbol{z}_n, 1),$$

where $\boldsymbol{x}_\circ = (3, 4)$ is the true parameter. Fig. 1.3(right) shows why averaging can be beneficial for SGD when the step size is not very small. Fig. 1.3(left) shows that, while gradient descent will outperform SGD when run for sufficiently long and with a fairly precise choice of step size, it is notably more

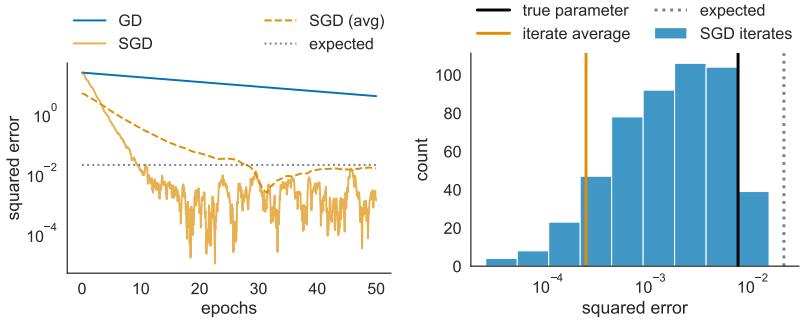
[2] As would usually be done in practice, the responses were centered to have mean zero, which in this case is equivalent to using the optimal intercept.



(a) $\eta_k = 1.0$



(b) $\eta_k = 0.1$



(c) $\eta_k = 0.01$

Figure 1.3: **(left)** Comparisons of gradient descent (GD) and stochastic gradient descent (SGD) with and without averaging. Computational effort is measured in epochs, where one epoch is equivalent to accessing N observations (so, 1 iteration of GD and N/B iterations of SGD). Squared error is the squared Euclidean between the estimate and the minimizer $\boldsymbol{x}_\star$. Both methods are run for 50 epochs and averaging is over the last 50% of iterates. **(right)** Histograms of the squared error between the last 50% of SGD iterates and $\boldsymbol{x}_\star$. In both plots, the expected squared error between the $\boldsymbol{x}_\star$ and true parameter is $2/N = 0.02$ while the actual squared error for true parameter (shown in the histogram plots) is approximately 0.007.

sensitive to the choice of η_k. Moreover, we expect the squared error $\|\boldsymbol{x}_\star - \boldsymbol{x}_\circ\|_2^2$ to be about $D/N = 0.02$ (for this dataset it is approximately 0.007). So for the purposes of either estimation or prediction, there is little benefit to approximating $\boldsymbol{x}_\star$ to squared error of less than about 10^{-4}. Thus, this example suggests that for many statistical and machine learning applications, SGD could provide sufficiently accurate solutions quite quickly.

While this example is suggestive of how SGD can work well, in practice obtaining good performance with SGD is more complicated. In particular, this book will address the following important questions:

Stochastic Optimization: Challenges and Questions

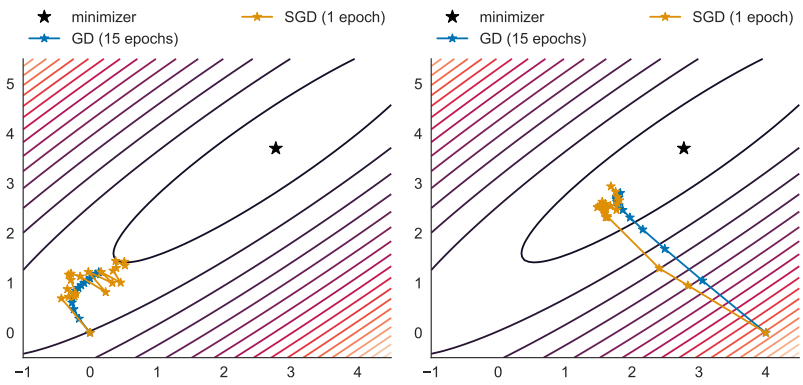
1. **General guarantees.** It is nice to see that empirically SGD works well on some specific losses and datasets. But to deploy a method, we would like general theoretical guarantees. So, in general, **When does SGD actually provide a good approximation to the minimizer? That is, under what conditions is SGD guaranteed to work?**

2. **Parameter tuning.** Usually we do not know the minimizer or true parameter – if we did, it would be unnecessary to use SGD. Without such knowledge, **How should the tuning parameters be set to achieve a desired behavior (e.g., level of accuracy)?** For SGD, these parameters are the batch size B, step size sequence $\{\eta_k\}_{k\in\mathbb{N}}$, and total number of iterations.

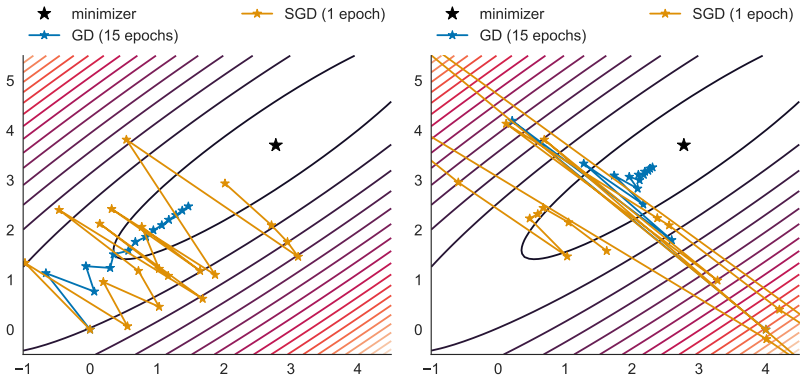
3. **Algorithm design.** SGD is a very simple – and maybe even a "naïve" – algorithm. **Can we design more computationally efficient algorithms that will work well on a wide variety of problems?** One approach is to consider fundamentally different algorithms. Or we might consider ***adaptive*** algorithms that adjust the SGD tuning parameters automatically.

To give an idea of why the last question is important, consider the following example where, because the components of $\boldsymbol{x}$ are strongly correlated (i.e., the problem is ill-conditioned), SGD can become unstable or converge very slowly.

Example 1.1.9 (GD and SGD performance on a loss function with elliptical contours)**.** *Consider the same setting as Example 1.1.8 but with a different*



(a) $\eta_k = 0.1$



(b) $\eta_k = 0.4$

Figure 1.4: Comparisons of gradient descent (GD) and stochastic gradient descent (SGD) optimization of the squared Euclidean loss function with correlated features. Computational effort is measured in epochs, as described in Example 1.1.8. The top row utilizes a fixed step size of 0.1, which proves adequate in the steeper regions (right column), but is much too small in the flatter regions (left column). The bottom row increases the fixed step size to 0.4, which results in the opposite problem.

data-generating process in which the components of $\boldsymbol{z}_n$ are correlated:

$$\boldsymbol{z}_n \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \qquad \textit{where} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}.$$

The highly correlated features result in a loss function with elliptical contours. Figure 1.4 shows how the gradient along the bottom-left to top-right diagonal is quite flat, while the gradient along the direction of the other diagonal is much steeper. For GD and SGD, this means that the ideal step size varies drastically depending on the current location in parameter space. For example, a larger step size is needed to prevent stalling in the regions with flatter gradient. However, a larger step size can cause oscillation or even divergent behavior in the steeper regions. Figure 1.4 demonstrates the dangers of naïvely choosing a fixed step size, with the resulting algorithms either stalling or experiencing oscillating behavior depending on the starting location.

1.2 Sampling

Other than optimization, perhaps the most common algorithmic challenge in data science is computing expectations with respect to a target distribution π defined on a parameter or state space $\mathcal{A}$. Specifically, given a vector-valued function $\boldsymbol{\phi} : \mathcal{A} \to \mathbb{R}^M$, we wish to calculate the expectation

$$\pi(\boldsymbol{\phi}) := \mathbb{E}_{\boldsymbol{X} \sim \pi}\{\boldsymbol{\phi}(\boldsymbol{X})\} = \int \boldsymbol{\phi}(\boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

For example, we might want to compute the mean $\boldsymbol{\mu}$ using $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$, the covariance using $\boldsymbol{\phi}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{x} - \boldsymbol{\mu})$, or a predictive p.d.f. at y using $\phi(\boldsymbol{x}) = p_{\boldsymbol{x}}(y)$ where $\{p_{\boldsymbol{x}}\}_{\boldsymbol{x} \in \mathcal{A}}$ is a family of predictive distributions. In the rest of this section, for simplicity we will focus on real-valued functions $\phi : \mathcal{A} \to \mathbb{R}$.

Most of the examples we will consider are from ***Bayesian statistics***. Assume we have observations $\mathcal{Y}$ and side information $\mathcal{Z}$ (such as covariates in a regression). We would like to learn a conditional model for $\mathcal{Y}$ given $\mathcal{Z}$. To do so, we choose a family of distributions called the ***observation model*** parameterized by $\boldsymbol{x} \in \mathcal{A}$, which we assume has a conditional p.d.f. $p(\mathcal{Y} \mid \boldsymbol{x}, \mathcal{Z})$. We must also choose a ***prior distribution*** for the parameters, which we also assume has a density $\pi_0(\boldsymbol{x})$. The prior distribution encodes any available information about what parameter values are *a priori* most

plausible. Together, they define a conditional joint distribution given by

$$g(\mathcal{Y}, \boldsymbol{x} \mid \mathcal{Z}) = p(\mathcal{Y} \mid \boldsymbol{x}, \mathcal{Z})\pi_0(\boldsymbol{x}).$$

Since we know $\mathcal{Y}$ and $\mathcal{Z}$, we can determine the conditional distribution of the parameters given $\mathcal{Y}$ and $\mathcal{Z}$, which is called the ***posterior distribution***. The posterior represents our updated beliefs about the most plausible parameter values given the observed data. Letting $p(\mathcal{Y} \mid \mathcal{Z}) := \int p(\mathcal{Y} \mid \boldsymbol{x}, \mathcal{Z})\pi_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ denote the ***marginal likelihood***, the posterior distribution has density

$$\pi(\boldsymbol{x} \mid \mathcal{Y}, \mathcal{Z}) = \frac{g(\mathcal{Y}, \boldsymbol{x} \mid \mathcal{Z})}{p(\mathcal{Y} \mid \mathcal{Z})} = \frac{p(\mathcal{Y} \mid \boldsymbol{x}, \mathcal{Z})\pi_0(\boldsymbol{x})}{p(\mathcal{Y} \mid \mathcal{Z})}.$$

So, in the Bayesian setting, we take $\pi(\boldsymbol{x}) = \pi(\boldsymbol{x} \mid \mathcal{Y}, \mathcal{Z})$

We first revisit the regression setting discussed in the previous section with responses $\mathcal{Y} = \{y_n\}_{n=1}^N$, where $y_n \in \mathbb{Y}$, and covariates $\mathcal{Z} = \{\boldsymbol{z}_n\}_{n=1}^N$, where $\boldsymbol{z}_n \in \mathbb{V}$.

Example 1.2.1 (Generalized linear models)**.** *Linear regression and logistic regression (Examples 1.1.1 and 1.1.2) are both* **generalized linear models** *(GLMs). In a GLM, the parameters are $\boldsymbol{x} = (\boldsymbol{\beta}, \boldsymbol{\psi})$, where $\boldsymbol{\beta} \in \mathbb{R}^D$ are the regression coefficients and $\boldsymbol{\psi} \in \Psi \subseteq \mathbb{R}^M$ are additional parameters. Observations are assumed independent and their response depends on the covariates only through the inner product $\boldsymbol{\beta}^\top \boldsymbol{z}$, so the observation model can be written as*

$$p(\mathcal{Y} \mid \boldsymbol{x}, \mathcal{Z}) = \prod_{n=1}^N p(y_n \mid \boldsymbol{\beta}^\top \boldsymbol{z}_n, \boldsymbol{\psi}).$$

Table 1.1 summarizes some common GLMs. Common prior distributions for the components of $\boldsymbol{\beta}$ include independent $\mathcal{N}(0, s^2)$ distributions or $\mathcal{T}(0, s^2, \nu)$ distributions, where the latter is a mean-zero t distribution with ν degrees of freedom. The linear function $\boldsymbol{\beta}^\top \boldsymbol{z}_n$ can also be replaced by a more complicated, nonlinear function such as a neural network (see Example 1.1.4). Placing a prior on the neural network parameters results in a **Bayesian neural network** *model.*

Next we revisit an unsupervised example from Section 1.1, where $\mathcal{Y} = \{\boldsymbol{y}_n\}_{n=1}^N$ and there are no covariates, so we drop dependence on $\mathcal{Z}$ from our notation.

Regression type	$\mathbb{Y}$	$\boldsymbol{\psi}$	observation model	$p(y \mid t, \boldsymbol{\psi})$
Gaussian linear	$\mathbb{R}$	$\sigma^2 > 0$	$y \mid \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\beta}^\top \boldsymbol{z}, \sigma^2)$	$\mathcal{N}(y \mid t, \sigma^2)$
Robust linear	$\mathbb{R}$	$s^2 > 0$, $\nu > 0$	$y \mid \boldsymbol{z} \sim \mathcal{T}(\boldsymbol{\beta}^\top \boldsymbol{z}, s^2, \nu)$	$\mathcal{T}(y \mid t, s^2, \nu)$
Logistic	$\{\pm 1\}$	$\emptyset$	$y \mid \boldsymbol{z} \sim \text{Bern}(\varphi(\boldsymbol{\beta}^\top \boldsymbol{z}))$	$\text{Bern}(y \mid \varphi(t))$
Poisson	$\mathbb{N}$	$\emptyset$	$y \mid \boldsymbol{z} \sim \text{Poiss}(\boldsymbol{\beta}^\top \boldsymbol{z})$	$\text{Poiss}(y \mid e^t)$

Table 1.1: **Common generalized linear models.** *Note:* $\varphi(t) = 1/(e^{-t}+1)$ denotes the logistic function.

Example 1.2.2 (Mixture models)**.** *Building on Example 1.1.5, the observation model for a Gaussian mixture model is given by*

$$p(\mathcal{Y} \mid \boldsymbol{x}) = \prod_{n=1}^{N} p_{\boldsymbol{x}}(\boldsymbol{y}_n),$$

where $p_{\boldsymbol{x}}$ is defined in Eq. (1.1). A Bayesian mixture model requires choosing a prior for the parameter vector $\boldsymbol{x}$. A common choice would be a factorized prior $\pi_0(\boldsymbol{x}) = \pi_0(\boldsymbol{w}) \prod_{k=1}^{K} \pi_0(\boldsymbol{\mu}_k)\pi_0(\boldsymbol{\Sigma}_k)$. For example, one might set $\pi_0(\boldsymbol{w}) = \text{Dir}(\boldsymbol{w} \mid \boldsymbol{\alpha})$, $\pi_0(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{0}, s^2\boldsymbol{I})$, and $\pi_0(\boldsymbol{\Sigma}) = \mathcal{W}(\boldsymbol{\Sigma} \mid \nu, \boldsymbol{\Sigma}_0)$, where Dir *denotes the Dirichlet distribution, $\mathcal{W}$ denotes the Wishart distribution, and the priors require the choice of hyperparameters $\boldsymbol{\alpha} \in \mathbb{R}_+^D$, $s^2 > 0$, $\nu \geq D - 1$, and $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{D \times D}$ positive semi-definite.*

In all these examples, the observations are assumed independent, so the log of the observation model density – often called the ***log likelihood*** when viewed as a function of the parameter $\boldsymbol{x}$ – can be written as

$$\log p(\mathcal{Y} \mid \boldsymbol{x}, \mathcal{Z}) = \sum_{n=1}^{N} \ell_{(n)}(\boldsymbol{x}),$$

where ℓ_n is the ***log likelihood*** associated with the nth observation in the dataset. In the regression setting, for example, $\ell_{(n)}(\boldsymbol{x}) = \log p(y_n \mid \boldsymbol{z}_n, \boldsymbol{x})$ while in the unsupervised setting $\ell_{(n)}(\boldsymbol{x}) = \log p_{\boldsymbol{x}}(\boldsymbol{y}_n)$.

As with optimization, there are special cases where the posterior distribution can be computed in closed form.[3] However, these solutions require very specific choices of prior distributions, which may be inappropriate. And, as

[3] See Section 7.6 of Murphy (2012) for the case of Bayesian linear regression.

with optimization, the computational cost often scales poorly with dimension. So, in general, we can usually only assume that the posterior density can be calculated up to the unknown marginal likelihood constant $p(\mathcal{D} \mid \mathcal{Z})$ through the pointwise evaluation of $g(\mathcal{Y}, \boldsymbol{x} \mid \mathcal{Z})$. While other approaches exist[4], we will focus on *sampling* methods that involve generating a set of samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ and approximating the expectation $\pi(\phi)$ by

$$\hat{\pi}_k(\phi) := \frac{1}{k} \sum_{\ell=1}^{k} \phi(\boldsymbol{x}_\ell).$$

In the data science context, the same four criteria we described for optimization algorithms apply (with some modification) when evaluating and designing sampling algorithms:

Sampling Algorithm Criteria

1. **General applicability.** New models are constantly being developed. Therefore, we will focus on general-purpose sampling algorithms, rather than specialized ones that exploit distinct problem structure.
2. **Dependence on dimensionality of $\boldsymbol{x}$.** Since $\boldsymbol{x}$ is often very high-dimensional, we want methods with computational cost that is linear or nearly linear in the dimension D. In some cases this isn't possible if, e.g., $g(\mathcal{Y}, \boldsymbol{x} \mid \mathcal{Z})$ requires $\Theta(D^2)$ operations to evaluate; in such cases we want the sampling algorithm to retain the same runtime dependence.
3. **Number of evaluations of $g(\mathcal{Y}, \boldsymbol{x} \mid \mathcal{Z})$.** When the dataset size is large or calculating $\mathcal{L}$ requires an expensive simulation (e.g., solving a system of ODEs or PDEs), it is important to compute $g(\mathcal{Y}, \boldsymbol{x} \mid \mathcal{Z})$ as few times as possible. In the sampling context we must also be careful about how the number of evaluations increases with the dimension.
4. **Obtaining a statistically accurate approximation.** Usually we would like $\hat{\pi}_k(\phi)$ to converge to $\pi(\phi)$ as $k \to \infty$. In practice we usually want the error of the approximation to be small relative to $\mathrm{Var}_\pi(\phi)$, the variance of $\phi(\boldsymbol{X})$ for $\boldsymbol{X} \sim \pi$.

We can summarize the sampling problem we wish to solve as follows:

[4] For example, see Chapters 21 and 22 of Murphy (2012)

Problem: Sampling with Unnormalized Densities

Design general-purpose sampling algorithms that, given a target distribution $\pi(\boldsymbol{x})$ that can only be evaluated pointwise up to a multiplicative constant, computes an estimate

$$\hat{\pi}_k(\phi) = \frac{1}{k}\sum_{\ell=1}^{k} \phi(\boldsymbol{x}_\ell) \approx \pi(\phi) = \mathbb{E}_{\boldsymbol{X}\sim\pi}\{\phi(\boldsymbol{X})\}$$

in a way that is scalable to high dimensions and large datasets while providing sufficient accuracy relative to $\mathrm{Var}_\pi(\phi)$.

When D is small, ***quasi-Monte Carlo*** and ***numerical quadrature*** methods can provide fast, accurate approximations of integrals. However, they are limited to low-dimensional settings because their approximation error grows exponentially in D (unless special problem structure can be exploited). A typical error bound for a k-sample approximation would be of order $\log(k)^{D+1}/k$. Thus, these methods tend to become impractical when $D \gtrsim 8$ and so fail to satisfy criterion 2.

If it is possible to generate independent samples from π, then ***simple Monte Carlo*** may be used instead. If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are i.i.d. samples from π, then by the law of large numbers,

$$\lim_{k\to\infty} \hat{\pi}_k(\phi) = \pi(\phi) \tag{1.6}$$

as long as $\mathbb{E}_{\boldsymbol{X}\sim\pi}\{|\phi(\boldsymbol{X})|\} < \infty$. Simple Monte Carlo is attractive since the estimate $\hat{\pi}_k(\phi)$ is unbiased and, if $\mathrm{Var}_\pi(\phi) < \infty$, the central limit theorem implies that

$$k^{1/2}\{\hat{\pi}_k(\phi) - \pi(\phi)\} \xrightarrow{d} \mathcal{N}(0, \sigma_\phi^2), \tag{1.7}$$

where $\xrightarrow{d}$ denotes convergence in distribution and $\sigma_\phi^2 := \mathrm{Var}_\pi(\phi)$. Therefore, the expected approximation error $\mathbb{E}\{|\hat{\pi}_k(\phi) - \pi(\phi)|\}$ is upper bounded by $\sigma_\phi/k^{1/2}$, so k need not be too large to ensure small error relative to the variance.

Hence, simple Monte Carlo satisfies criterion 4. Moreover, since the error is independent of the dimension, it is effective in high-dimensional problems, thereby satisfying criterion 2. However, in most problems of interest it is impossible to generate independent samples, so simple Monte Carlo fails to satisfy criterion 1.

1.2.1 Simple Monte Carlo

While simple Monte Carlo has limitations, it remains an extremely value method that provides an important baseline when comparing to alternatives. Let $\hat{\pi}_k(\phi)$ be the simple Monte Carlo estimator of $\pi(\phi)$, the empirical mean of samples $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \overset{\text{iid}}{\sim} \pi$. We will examine the error between the random variable $\hat{\pi}_k(\phi)$ and the true expectation $\pi(\phi)$. First note that $\hat{\pi}_k(\phi)$ is an unbiased estimator since

$$\mathbb{E}\{\hat{\pi}_k(\phi)\} = \frac{1}{k}\sum_{\ell=1}^{k}\mathbb{E}\{\phi(\boldsymbol{x}_\ell)\} = \frac{1}{k}\sum_{\ell=1}^{k}\pi(\phi) = \pi(\phi),$$

following from the linearity of expectation and the assumption $\boldsymbol{x}_\ell \sim \pi$, but not requiring the independence assumption. The Monte Carlo error in this context is defined as the expected squared difference between $\hat{\pi}_k(\phi)$ and $\pi(\phi)$, which is equivalent to the variance of the estimator

$$\mathbb{E}\{[\hat{\pi}_k(\phi) - \pi(\phi)]^2\} = \mathrm{Var}\{\hat{\pi}_k(\phi)\},$$

following from the unbiasedness of $\hat{\pi}_k(\phi)$. Since the summands $\phi(\boldsymbol{x}_\ell)$ are independent, the variance operator can be passed inside the summation to obtain

$$\mathrm{Var}\{\hat{\pi}_k(\phi)\} = \frac{1}{k^2}\sum_{\ell=1}^{k}\mathrm{Var}\{\phi(\boldsymbol{x}_\ell)\} = \frac{1}{k^2}\sum_{\ell=1}^{k}\sigma_\phi^2 = \frac{\sigma_\phi^2}{k}. \tag{1.8}$$

We have established that the expected squared error of a simple Monte Carlo estimate decays at rate k^{-1}.[5] Since this is on the squared scale, one commonly hears that the simple Monte Carlo error rate is $k^{-1/2}$. Importantly, this rate is independent of the dimension D.[6] While unbiasedness and the above variance expression hold for any finite k, Eq. (1.6) provides the important asymptotic justification that the simple Monte Carlo estimate approaches the true expectation as $k \to \infty$. In practice, we utilize a finite set of k samples and thus it is essential to quantify the error in the estimate. Eq. (1.8) provides a useful start, giving a notion of how "spread out" the error may

[5] The expression for the simple Monte Carlo variance reveals that there are two options for reducing the error: decreasing σ_ϕ^2 or increasing the number of samples k. There are many so-called **variance reduction methods** that seek to accomplish the former.

[6] While it is true that the error rate is independent of D, it is misleading to say that simple Monte Carlo methods are immune to the curse of dimensionality. Often, the cost of drawing i.i.d. samples from π scales poorly with D.

be, but provides no information on the shape of the error distribution. For large enough k, the central limit theorem Eq. (1.7) provides this information, allowing for the construction of (approximate) confidence intervals. Indeed, the central limit theorem

$$\frac{k^{1/2}}{\sigma_\phi}\{\hat{\pi}_k(\phi) - \pi(\phi)\} \xrightarrow{d} \mathcal{N}(0,1)$$

implies convergence of distribution functions. Thus, letting $Z \sim \mathcal{N}(0,1)$ and $z_\alpha \in \mathbb{R}$ we have

$$\mathbb{P}\left\{-z_\alpha \leq \frac{\hat{\pi}_k(\phi) - \pi(\phi)}{\sigma_\phi / k^{1/2}} \leq z_\alpha\right\} \to \mathbb{P}\{-z_\alpha \leq Z \leq z_\alpha\}.$$

For a $100(1-\alpha)\%$ confidence interval, choose z_α such that $\mathbb{P}\{-z_\alpha \leq Z \leq z_\alpha\} = 1-\alpha$. Then we obtain

$$\mathbb{P}\left\{\hat{\pi}_k(\phi) - \frac{\sigma_\phi z_\alpha}{k^{1/2}} \leq \pi(\phi) \leq \hat{\pi}_k(\phi) + \frac{\sigma_\phi z_\alpha}{k^{1/2}}\right\} \to 1-\alpha. \tag{1.9}$$

For finite k, the interval $\left[\hat{\pi}_k(\phi) - \frac{\sigma_\phi z_\alpha}{k^{1/2}}, \hat{\pi}_k(\phi) + \frac{\sigma_\phi z_\alpha}{k^{1/2}}\right]$ is thus an approximate $100(1-\alpha)\%$ confidence interval.[7] Note that σ_ϕ is typically unknown, so to construct this confidence interval in practice it is common to replace σ_ϕ with its empirical estimate.

Example 1.2.3 (Simple Monte Carlo Error Analysis)**.** *Figure 1.5 illustrates how the error in simple Monte Carlo estimates change with sample size. The sampling distribution of $\hat{\pi}_k(\phi)$ is explored by independently replicating the experiment many times for varying sample sizes k.*

1.2.2 Markov chain Monte Carlo

Since simple Monte Carlo is usually not possible, instead of attempting to generate perfect samples (that is, samples with exactly the distribution

[7] It is important to recall the proper interpretation of confidence intervals. The bounds of the confidence interval are the random variables under consideration here, while the value $\pi(\phi)$ is fixed. Thus, the statement Eq. (1.9) should be interpreted as follows. Suppose we sample $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \overset{\text{iid}}{\sim} \pi$, form the estimate $\hat{\pi}_k(\phi)$, and construct the associated $100(1-\alpha)\%$ confidence interval. Now repeat this experiment independently many times, constructing a confidence interval each time. Then approximately $100(1-\alpha)\%$ of these intervals will contain the true expectation $\pi(\phi)$.

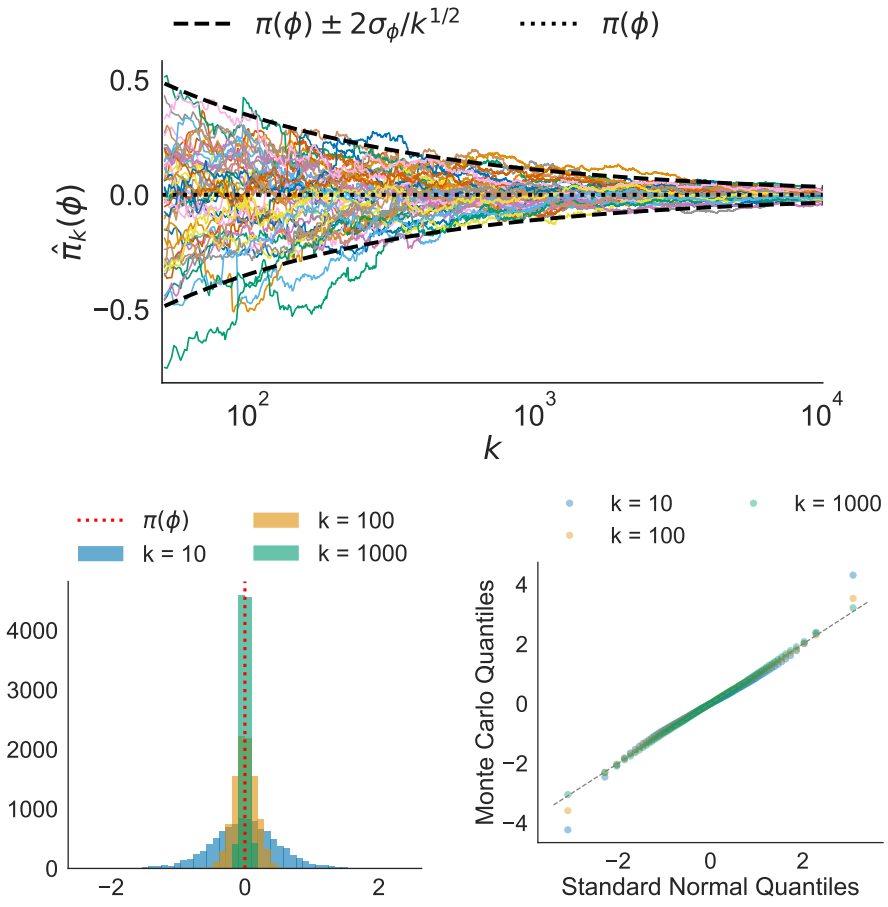


Figure 1.5: Simple Monte Carlo error results summarized over an ensemble of independent replicate experiments. The target distribution π is a t distribution with 3 degrees of freedom. We consider ϕ to be the identity so that $\pi(\phi)$ is simply the mean of the t distribution, which is equal to 0. For each sample size k, 10,000 experiments were conducted, resulting in 10,000 samples from the sampling distribution of $\hat{\pi}_k(\phi)$. **(top)** Samples from the sampling distribution of $\hat{\pi}_k(\phi)$, demonstrating the convergence of the estimate as $k \to \infty$. The plotted error bars $\pi(\phi) \pm \frac{2\sigma_\phi}{k^{1/2}}$ help to demonstrate the $k^{-1/2}$ convergence rate. **(bottom left)** Histograms of samples of $\hat{\pi}_k(\phi)$ for different samples sizes k. The decreasing spread of the sampling distribution as k increases provides another demonstration of convergence, while the approximately Gaussian shape of the distributions illustrates the predictions of the central limit theorem. **(bottom right)** An alternative view of the central limit theorem in action. The theoretical quantiles of the standard normal distribution are plotted against the empirical quantiles of the normalized sampling distributions of $\hat{\pi}_k(\phi)$ at various sample sizes. Even small sample sizes show reasonable agreement between the distributions, but larger sample sizes are required for the tails of the distributions to agree.

π), we can instead aim for a less ambitious goal similar in spirit to the iterative optimization approach of repeatedly improving the quality of an approximation. Specifically, we generate a random sequence of iterates $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ such that the *distribution* of the iterates gets closer and closer to π. This is the key idea of ***Markov chain Monte Carlo (MCMC)***, which is a flexible and broadly applicable alternative to simple Monte Carlo. In its simplest form, given all previous samples, the distribution of $\boldsymbol{x}_{k+1}$ depends only on $\boldsymbol{x}_k$. Thus, the behavior of the MCMC algorithm depends on just the initial distribution of $\boldsymbol{x}_0$ and the family of conditional distributions $q(\boldsymbol{x} \mid \boldsymbol{x}')$. As in stochastic optimization, we can discard the early iterates, which we expect to have a distribution far from π. For example, paralleling Eq. (1.5), we can use the most recent 50% of iterations to construct the estimator

$$\hat{\pi}_k(\phi) := \frac{1}{\lceil k/2 + 1 \rceil} \sum_{\ell=\lfloor k/2 \rfloor}^{k} \phi(\boldsymbol{x}_\ell).$$

While both simple Monte Carlo and MCMC estimate expectations via empirical averages, it is important to keep in mind the fundamental differences between the two approaches. Simple Monte Carlo forms estimates using i.i.d. samples. But MCMC samples are neither independent nor identically distributed. This leads to various practical and theoretical challenges, but the applicability of MCMC in a wide array of settings where simple Monte Carlo is infeasible justifies these additional challenges. In Example 1.2.3 we examined the Monte Carlo error when the estimator $\hat{\pi}_k(\phi)$ was computed using i.i.d. samples. We now consider the same estimator but constructed using samples $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ produced via MCMC. The first issue to tackle is that the samples are no longer identically distributed according to π. Rather, the marginal distribution of $\boldsymbol{x}_\ell$ approaches π as $\ell \to \infty$. Thus, while the MCMC estimate $\hat{\pi}_k(\phi)$ may be biased for finite sample size k, this bias vanishes as $k \to \infty$. In practice, there are a variety of diagnostics to assess whether the chain has converged to the target distribution. When constructing the estimator $\hat{\pi}_k(\phi)$, we exclude these early samples deemed to have been obtained prior to convergence. This early portion of the iterates is often referred to as the **burn-in**.

Yet even in the idealized scenario where $\boldsymbol{x}_0 \sim \pi$ (perhaps after re-indexing to zero after dropping burn-in samples) so the estimate is unbiased, the samples $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are still correlated. This lack of independence prevents exchanging the order of the variance and summation as in the simple Monte

Carlo variance calculation. Instead, we obtain

$$\begin{aligned}\mathrm{Var}\{\hat{\pi}_k(\phi)\} &= \frac{1}{k^2}\,\mathrm{Var}\left\{\sum_{\ell=1}^{k}\phi(\boldsymbol{x}_\ell)\right\} \\ &= \frac{1}{k^2}\sum_{\ell=1}^{k}\mathrm{Var}\{\phi(\boldsymbol{x}_\ell)\} + \frac{2}{k^2}\sum_{\ell>\ell'}\mathrm{Cov}\{\phi(\boldsymbol{x}_\ell),\phi(\boldsymbol{x}_{\ell'})\} \\ &= \frac{\sigma_\phi^2}{k} + 2\sum_{\ell=1}^{k-1}\frac{k-\ell}{\ell}\,\mathrm{Cov}\{\phi(\boldsymbol{x}_0),\phi(\boldsymbol{x}_\ell)\}\end{aligned}$$

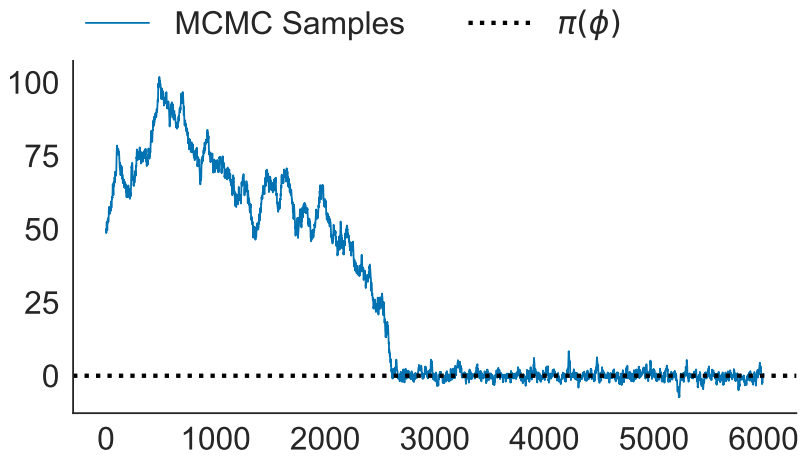
where the final inequality uses the assumption $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \sim \pi$. We notice the first term is precisely the simple Monte Carlo error given in Eq. (1.8), while the second term accounts for correlation in the samples. In theory, this correlation could be negative and hence result in smaller error. However, in practice this term is almost always positive and hence represents a penalty on top of the typical simple Monte Carlo error. This is the price to be paid for using correlated samples, which carry less information than an equal number of independent samples. Therefore, common diagnostic tests for MCMC involve investigating the ***autocorrelation*** $\mathrm{Cov}\{\phi(\boldsymbol{x}_0),\phi(\boldsymbol{x}_\ell)\}/\sigma_\phi^2$ for different values of the lag ℓ. Autocorrelation that decays very slowly as ℓ increases is a reason for concern. In practice, $\mathrm{Cov}\{\phi(\boldsymbol{x}_0),\phi(\boldsymbol{x}_\ell)\}$ is not known so must be estimated empirically using the MCMC samples.

Example 1.2.4 (Comparing Simple Monte Carlo to MCMC)**.** *Figures 1.6 and 1.7 compare MCMC to simple Monte Carlo, continuing from the example presented in Example 1.2.3 of estimating the mean of a t distribution with* 3 *degrees of freedom. While MCMC requires more samples to obtain accurate estimates of expectations, it still provides a good approximate after removing the burn-in samples.*

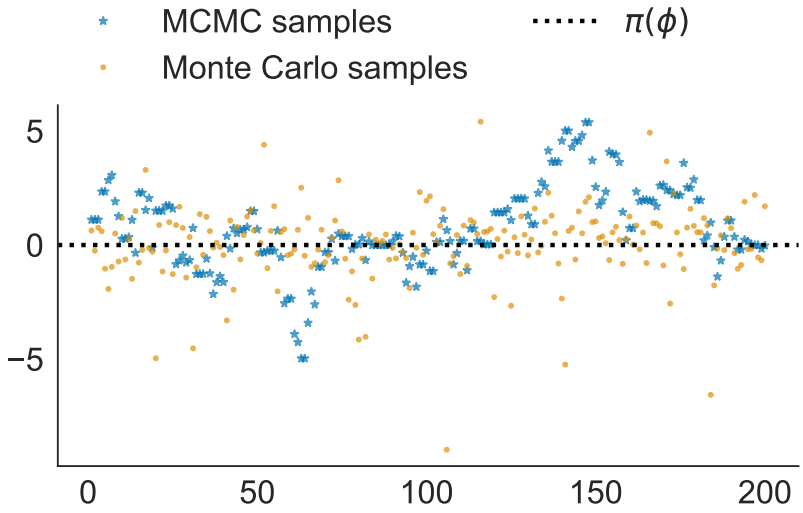
While this example shows the potential utility of MCMC, is also hints at some of the complications with using it in practice and analyzing the convergence and approximation properties of MCMC algorithms.

MCMC: Challenges and Questions

1. **General guarantees.** As with SGD, it is good to see MCMC seems to work on some specific problems. But we would like general theoretical guarantees so we can have confidence MCMC will work on new problems. In particular, **When does an MCMC algorithm**



(a) All MCMC samples.



(b) First 200 MCMC samples after burn-in compared to a sequence of i.i.d. Monte Carlo samples.

Figure 1.6: MCMC versus Monte Carlo samples. The setup is the same as Example 1.2.3, with target distribution π a t distribution with 3 degrees of freedom and mean 0. However, we consider here only a single experiment, as opposed to the previous example which analyzed an ensemble of independent replicates. **(top)** The complete set of MCMC samples. The Markov chain was initialized in a location far out in the tails of π. The chain takes a few thousand iterations to reach the region of high probability. This initial set of samples is typically dropped in order to reduce the error in the estimate $\hat{\pi}_k(\phi)$. While it visually appears that dropping the first 3,000 samples would be sufficient, in the following plots we conservatively drop the first 10,000 samples as burn-in. **(bottom)** The first 200 samples retained after dropping the burn-in samples compared to i.i.d. samples. By zooming in, the correlation between the MCMC samples compared to the lack of correlation in the i.i.d. samples is more noticeable.

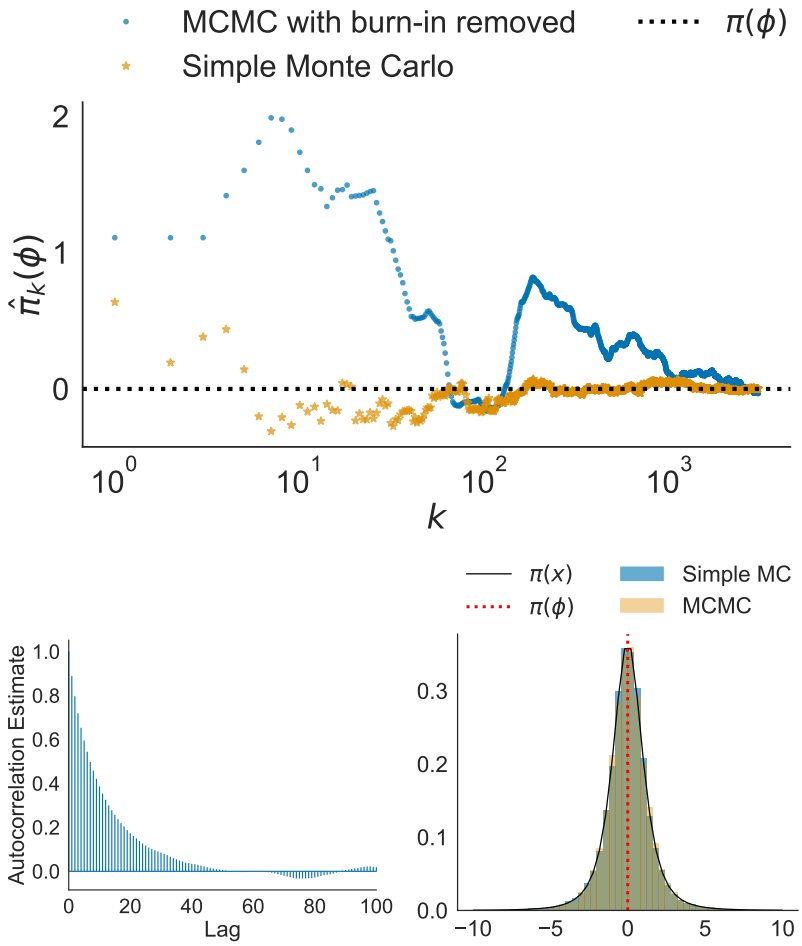


Figure 1.7: Comparing simple Monte Carlo with MCMC. The setup is the same as Fig. 1.6. All plots are constructed by excluded the first 10,000 MCMC samples, which are dropped as burn-in. (**top**) Comparing the convergence of the estimate $\hat{\pi}_k(\phi)$ for simple Monte Carlo and MCMC. **(bottom left)** Estimates of the MCMC autocorrelation $\text{Cov}\{\phi(\boldsymbol{x}_0), \phi(\boldsymbol{x}_\ell)\}/\sigma^2_\phi$ at different lags ℓ. **(bottom right)** Histograms of the simple Monte Carlo and MCMC samples compared to the density $\pi(x)$.

satisfy a law of large numbers and a central limit theorem? Because the samples we use are dependent, these questions are significantly more challenging to answer than for simple Monte Carlo where the samples are independent.

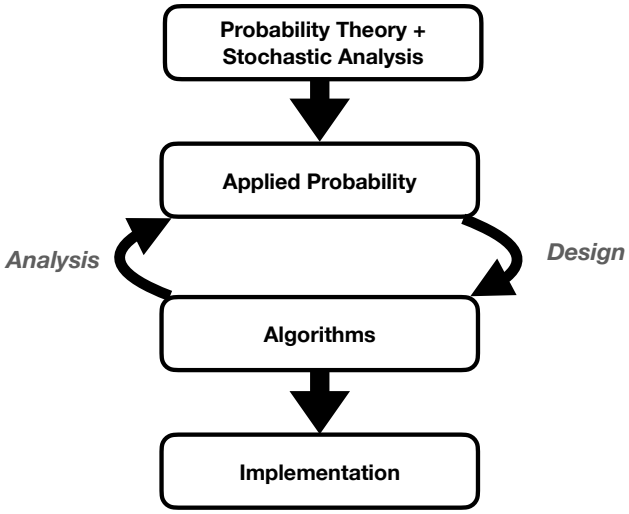
2. **Parameter tuning.** Usually a given MCMC procedure has a variety of tuning parameters (number of iterations, step size, etc.). **How should the tuning parameters be set to achieve a desired level of approximation accuracy for expectation estimates in the most computationally efficient manner?**

3. **Algorithm design.** There are innumerable MCMC algorithms that could be used to sample from a particular target distribution. But some algorithms may be orders of magnitude more efficient. **How do we design computationally efficient algorithms that will work well on a wide variety of problems?** As with stochastic optimization, we might consider ***adaptive*** algorithms that adjust the MCMC tuning parameters automatically.

1.3 Stochastic Methods

SGD and MCMC are both examples of ***randomized iterative algorithms***. By *iterative* we mean they repeatedly update some state $\boldsymbol{x}$ over and over again, only stopping after a fixed number of iterations or after some stopping criterion is met. They are *randomized (a.k.a. stochastic)* in the sense that each iteration used some external source of randomness (or pseudorandomness). Hence, each time the algorithm is run the result is different. These features means we can view both types of algorithms as *stochastic processes*. Hence, we can use the theory of stochastic processes to analyze algorithm behavior and what effect various tuning parameters have on algorithm performance. We can also use stochastic process theory to design new algorithms. For example, we can construct an "ideal" stochastic process, then numerically approximate the process to arrive at an implementable algorithm. It will often be fruitful to analyze such an algorithm by comparing it to the originating stochastic process or some other "simpler" process that's easier to understand. In short, the themes of the class will be to (1) start with a stochastic process and use it to design an algorithm, and (2) start with an algorithm, and use stochastic process methods to analyze it. As a practical matter, this will require us to use tools from probability theory, of which stochastic processes

is an important subfield, and stochastic analysis, which combines ideas from probability theory and calculus. We refer to all these mathematical tools together as ***stochastic methods***: hence the title of the book. Sometimes these tools will be quite advanced. So, rather than developing them from the bottom up – as you might do in a typical probability theory or stochastic processes course – we will mostly take them as given, then apply them to design and analyze algorithms. As such you might think about this course as being a mix of "applied probability" and "analysis of algorithms":



Chapter 2

Probability Theory

A concise review of probability theory using measures – but without measure theory. The focus is on intuitions and important results rather than proofs and careful treatment of technicalities.

Before diving into stochastic process theory and other stochastic methods, we review the necessary probability theory background. We operate at a level of technical sophistication greater than what is seen in the typical introductory undergraduate probability theory courses but short of a graduate-level, measure-theoretic treatment.

2.1 Events and Probabilities

The usual starting point for probability theory is to define a set of possible outcomes for an "experiment," which should be interpreted in the broadest terms possible. For example, the temperature tomorrow at noon, the outcome of a roll of a dice, the card I select at random from a deck, or the distance I bike today are all experiments. In all these cases, there are many possible outcomes.

Definition 2.1.1 (Sample space)**.** *A set of all possible outcomes of an experiment is called the* **sample space**.

We will follow tradition and denote the sample space by Ω.

Example 2.1.2 (Coin flipping)**.** *If we flip a single coin, there are two possible outcomes, heads or tails, which we denote as* H *and* T*. Thus,* $\Omega = \{\mathsf{H}, \mathsf{T}\}$*. If*

instead we flip 3 coins in a row, the sample space is all 8 possible combinations of three heads and tails: $\Omega = \{\mathsf{HHH}, \mathsf{HHT}, \mathsf{HTH}, \mathsf{HTT}, \mathsf{THH}, \mathsf{THT}, \mathsf{TTH}, \mathsf{TTT}\}$.

Example 2.1.3 (Selecting a card)**.** *If we select a card at random from a standard 52-card deck,* Ω *has 52 elements, one for each card. Recall that a card has one of 4 possible suits (heart, diamonds, spades, clubs) and and 13 possibles ranks (ace, 2 through 10, jack, queen, king).*

In the coin flipping example, we might be interested in whether (a) there are no heads or (b) whether there are at least two tails. Each of these can be represented as a set of possible outcomes S: (a) $S = \{\mathsf{TTT}\}$ and (b) $S = \{\mathsf{HTT}, \mathsf{THT}, \mathsf{TTH}, \mathsf{TTT}\}$. A set $S \subseteq \Omega$ is called an ***event***. We will write $\mathbb{P}(S)$ to denote the probability of the event S. More precisely, for an arbitrary set S, define the ***power set*** $\mathcal{P}(S) := \{S' : S' \subseteq S\}$, the set of all subsets of S (noting that S is a subset of itself). Hence, $\mathbb{P}$ is in fact a function $\mathbb{P}\colon \mathcal{P}(\Omega) \to [0, 1]$.

We would like to define the probability of events in a self-consistent way, which leads to the ***two axioms of probability***. The first is based on the idea that some outcome in the sample space must happen, so

(Axiom 1) $$\mathbb{P}(\Omega) = 1.$$

The second axiom relates the probabilities of different events. It is motivated by the observation that if $S, S' \subseteq \Omega$ are disjoint (that is, $S \cap S' = \emptyset$), then the event S might happen or the event S' might happen, but both events cannot happen. Therefore, probability of $S \cup S'$ should be equal to the sum of the respective probabilities of S and S'. To allow for the possibility of considering more than two events, instead consider ***pairwise disjoint*** events $S_1, S_2, \cdots \in \mathcal{P}(\Omega)$, meaning that $S_i \cap S_j = \emptyset$ if $i \neq j$. We require that

(Axiom 2) for pairwise disjoint $S_1, S_2, \cdots \in \mathcal{P}(\Omega)$, $\mathbb{P}(\cup_{i \geq 1} S_i) = \sum_{i \geq 1} \mathbb{P}(S_i)$.

The axioms imply results like $\mathbb{P}(\emptyset) = 0$ which match with the intuition that the probability that nothing happens should be zero.

Lemma 2.1.4. *The probability of the null event is zero:* $\mathbb{P}(\emptyset) = 0$.

Proof. Assume $\mathbb{P}(\emptyset) = a > 0$. Then using Axiom 2 we have $a = \mathbb{P}(\emptyset) = \mathbb{P}(\emptyset \cup \emptyset) = \mathbb{P}(\emptyset) \cup \mathbb{P}(\emptyset) = a + a = 2a$. But if $2a = a$, then $a = 0$, a contradiction. □

Exercise 2.1.1 (Probability formulas)**.** *Show that (a) for any* $S \in \mathcal{P}(\Omega)$*,* $\mathbb{P}(S^c) = 1 - \mathbb{P}(S)$ *and (b) for any* $S, S' \in \mathcal{P}(\Omega)$*,* $\mathbb{P}(S \cup S') = \mathbb{P}(S) + \mathbb{P}(S') - \mathbb{P}(S \cap S')$*.*

2.2 Random Variables

After defining events, the next step is to define a ***random variable*** $X \colon \Omega \to \mathbb{R}$. We can then define events related to the value of the random variable. For example, the event that X is between a and b would be

$$S = \{\omega \in \Omega \,:\, X(\omega) \in [a, b]\}.$$

More generally, the event that $X \in A \subseteq \mathbb{R}$ can be written using the function inverse notation

$$X^{-1}(A) := \{\omega \in \Omega \,:\, X(\omega) \in A\}.$$

The probability of this event is $\mathbb{P}(X^{-1}(A)) =: \mathbb{P}\{X \in A\}$.

Introductory probability theory courses usually introduce two kinds of random variables: discrete and continuous. Each kind is based on a different assumption about the form of $\mathbb{P}\{X \in A\}$. Given a function $\phi \colon \mathbb{R} \to \mathbb{R}$, this special form is then used to define the *expectation* $\mathbb{E}\{\phi(X)\}$, which informally is the value that $\phi(X)$ takes "on average." The definitions used in the discrete and continuous cases are reasonable but also *ad hoc*.

2.2.1 Discrete Random Variables

A ***discrete random variable*** $Y \colon \Omega \to \mathbb{R}$ takes on only a finite or countable number of values V and is fully determined by its ***probability mass function*** (p.m.f.) $p_Y \colon V \to [0, 1]$, which must satisfy $\sum_{y \in V} p_Y(y) = 1$. In particular, we set $\mathbb{P}\{Y = y\} = p_Y(y)$. Using Axiom 2 we can determine the probability of any event A as

$$\mathbb{P}\{Y \in A\} = \sum_{y \in A \cap V} p_Y(y). \tag{2.1}$$

The expectation of $\phi(Y)$ is defined as

$$\mathbb{E}\{\phi(Y)\} := \sum_{y \in V} \phi(y) p_Y(y). \tag{2.2}$$

Example 2.2.1. *The* **Bernoulli distribution** $\mathrm{Bern}(q)$ *with parameter* $q \in [0,1]$ *has the p.m.f.* $p_{\mathrm{Bern}(q)} \colon y \mapsto q^y(1-q)^{1-y}$*, where* $x \in \{0,1\}$*. We write* $Y \sim \mathrm{Bern}(q)$ *to denote that* Y *has a Bernoulli distribution. The expected value is given by* $\mathbb{E}\{\phi(Y)\} = \phi(0)(1-q) + \phi(1)q$*.*

Example 2.2.2. *The* **binomial distribution** $\mathrm{Binom}(n,q)$ *with parameters* $n \in \mathbb{N}$ *and* $q \in [0,1]$ *has p.m.f.* $p_{\mathrm{Binom}(n,q)} \colon y \mapsto \binom{n}{y} q^y(1-q)^{n-y}$*, where* $y \in \{0,\dots,n\}$*. For a binomial random variable* $Y \sim \mathrm{Binom}(n,p)$ *and* $B \subseteq \mathbb{R}$*,*

$$\mathbb{P}(Y \in B) = \sum_{y \in B \cap \{0,\dots,n\}} \binom{n}{y} q^y (1-q)^{n-y}.$$

Note that $\mathrm{Binom}(1,q) = \mathrm{Bern}(q)$*. We write* $Y \sim \mathrm{Binom}(n,q)$ *to denote that* Y *has a binomial distribution. The expected value is given by* $\mathbb{E}\{\phi(Y)\} = \sum_{y=0}^{n} \phi(y)\binom{n}{y} q^y(1-q)^{n-y}$*.*

2.2.2 Continuous Random Variables

A ***continuous random variable*** X on $\mathbb{R} \colon \Omega \to \mathbb{R}$ is determined by its ***probability density function*** $f_X \colon \mathbb{R} \to [0,1]$ that satisfies $\int f_X(x)\mathrm{d}x = 1$. The probability of $X \in A$ is defined as

$$\mathbb{P}\{X \in A\} := \int_A f_X(x)\mathrm{d}x.$$

The expectation of $\phi(X)$ is defined as

$$\mathbb{E}\{\phi(X)\} := \int \phi(x) f_X(x)\mathrm{d}x. \tag{2.3}$$

The definition is similar to the discrete case but using an integral, which we can think of as the continuous equivalent of a sum.

Example 2.2.3. *The* **Gaussian distribution** $\mathcal{N}(\mu, \sigma^2)$ *with parameters* $\mu \in \mathbb{R}$ *and* $\sigma^2 > 0$ *has density*

$$f_{\mathcal{N}(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Of course for $X \sim \mathcal{N}(\mu,\sigma^2)$*, the mean is equal to* $\mathbb{E}(X) = \mu$ *and the variance is* $\mathbb{E}\{(X-\mu)^2\} = \sigma^2$*.*

Example 2.2.4. *Define the* **indicator function** $\mathbb{1}(C)$ *to be equal to 1 if condition C is true and equal 0 if C is false. The* **exponential distribution** $\text{Exp}(\lambda)$ *with* **rate parameter** $\lambda > 0$ *has density*

$$f_{\text{Exp}(\lambda)}(x) = \lambda e^{-\lambda x} \mathbb{1}(x \geq 0).$$

For $X \sim \text{Exp}(\lambda)$, the mean is equal to $\mathbb{E}(X) = \lambda^{-1}$ and the variance is $\mathbb{E}\{(X - \lambda^{-1})^2\} = \lambda^{-2}$.

2.3 A Unified Approach to Random Variables

Having separate definitions for discrete and continuous random variables turns out to be rather awkward.

Example 2.3.1. *Consider the distance I bike tomorrow as a random variable X. There is usually some positive probability I will not bike, in which case the distance I bike is zero. But if I do bike, the distance I bike is best captured by a continuous distribution. However, with the tools from introductory probability, I cannot define such a "biking distance" random variable since it is neither discrete nor continuous, but a mix of the two.*

Situations with such "mixed" random variabels are quite common. For example, the random variable for the amount of rain tomorrow would have a similar behavior: there is some positive probability of no rain; but if there is rain, the amount of rain has a continuous distribution.

It turns out there is a solution to this problem that may appear simplistic but turns out to be a very powerful approach: rather than think of a random variable X as being defined by its p.m.f. or p.d.f., we can simply think of it being determined by the values of $\mathbb{P}\{X \in A\}$ for all possible $A \in \mathcal{P}(\mathbb{R})$.[1] We can summarize these probabilities by defining a function $\mu \colon \mathcal{P}(\mathbb{R}) \to [0, 1]$ given by

$$\mu(A) := \mathbb{P}\{X \in A\}.$$

Such functions satisfy the same axioms as $\mathbb{P}$ does (see Lemma 2.3.4 below), and are generically referred to as ***probability measures***.

[1] Technically, we need to limit ourselves to subsets that belong to an appropriate σ-algebra. But we will ignore these (important) details because any reasonable set you can think of will be allowed.

Definition 2.3.2 (Probability measure)**.** *Given a sample space Ω, a function $\mu\colon \mathcal{P}(\Omega) \to [0,1]$ is called a* **probability measure** *(or* **distribution***) if* ***(a)*** *$\mu(\Omega) = 1$ and* ***(b)*** *for pairwise disjoint $S_1, S_2, \dots \in \mathcal{P}(\Omega)$,*

$$\mu(\cup_{i\geq 1} S_i) = \sum_{i\geq 1} \mu(S_i).$$

The pair (Ω, μ) is called a **probability space***.*[2]

Notice that the definition of a probability measure does not require any particular choice of Ω. For example, we could have $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^D$. But we could also choose $\Omega = \mathbb{Z} \times \mathbb{R}$ or $\Omega = \{\mathsf{H}, \mathsf{T}\}$. This generality motivates the following definitions and naming conventions.

Definition 2.3.3 (Random variables and distributions)**.** *Given a probability space $(\Omega, \mathbb{P})$ and set $\mathcal{A}$, a* **random element** *(or* **random variable***) is a function $X\colon \Omega \to \mathcal{A}$. The probability that X takes on a value in a set $A \in \mathcal{P}(\mathcal{A})$ is $\mathbb{P}\{X \in A\} := \mathbb{P}(\{\omega \in \Omega \,:\, X(\omega) \in A\})$. The* **distribution** *(or* **law***) of a random variable is the probability measure $\mathcal{L}_X\colon \mathcal{P}(\mathcal{A}) \to [0,1]$ defined as $\mathcal{L}_X(A) := \mathbb{P}\{X \in A\}$ for all $A \in \mathcal{P}(\mathcal{A})$. If $\mu = \mathcal{L}_X$, then we write $X \sim \mu$ (read: "X has distribution μ").*

Lemma 2.3.4. *The law of a random variable is a valid probability measure.*

Proof. We prove that Definition 2.3.2(a) holds and leave verification of Definition 2.3.2(b) as an exercise. Since $X(\omega) \in \mathcal{A}$ for all $\omega \in \Omega$ it follows that

$$\{\omega \in \Omega \,:\, X(\omega) \in \mathcal{A}\} = \Omega.$$

By the definition of $\mathcal{L}_X$ and the assumption that $\mathbb{P}$ is a probability measure, we therefore have that

$$\begin{aligned}\mathcal{L}_X(\mathcal{A}) &= \mathbb{P}\{X \in \mathcal{A}\} \\ &= \mathbb{P}(\{\omega \in \Omega \,:\, X(\omega) \in \mathcal{A}\}) \\ &= \mathbb{P}(\Omega) \\ &= 1.\end{aligned}$$

□

[2] Technically, a probability space also requires a choice of σ-algebra. See **??** if you are interested in the details.

Exercise 2.3.1. *Finish the proof of Lemma 2.3.4 by showing that $\mathcal{L}_X$ satisfies Definition 2.3.2(b).*

In general, when defining random variables, we will always have a "background" probability space $(\Omega, \mathbb{P})$ in mind but we will rarely specify it explicitly. This background probability space ensures a precise meaning to statements about one or more random variables. For example if $X\colon \Omega \to \mathbb{R}$ and $Y\colon \Omega \to \mathbb{R}$ are real-valued random variables, and $A, B \subseteq \mathbb{R}$, then the expression $\mathbb{P}\{X \in A, Y \in B\}$ means "the probability of the event where $X \in A$ and $Y \in B$." The probability measure we are using here is $\mathbb{P}$, so the event is in the background probability space:

$$\mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}(\{\omega \in \Omega \,:\, X(\omega) \in A \text{ and } Y(\omega) \in B\}).$$

Example 2.3.5. *Let $\Omega = \{0, 1\}$ and $\mathbb{P}(\{0\}) = \mathbb{P}(\{1\}) = 1/2$. If $\mathcal{A} = \mathbb{R}$ and $X(\omega) = 4\omega$, then*

$$\begin{aligned}\mathbb{P}(X = 0) &= \mathbb{P}(\{\omega \in \{0, 1\} \,:\, X(\omega) = 0\}) = \mathbb{P}\{0\} = 1/2\\ \mathbb{P}(X = 4) &= \mathbb{P}(\{\omega \in \{0, 1\} \,:\, X(\omega) = 4\}) = \mathbb{P}\{1\} = 1/2.\end{aligned}$$

So $\mathbb{P}\{X \in S\} = 0.5\mathbb{1}(0 \in S) + 0.5\mathbb{1}(4 \in S)$ and hence X is uniformly distributed on the set $\{0, 4\}$.

Example 2.3.6. *Let X denote the distance I will bike tomorrow. Let p be the probability I do not bike. If I do bike, assume the distribution of how far I bike has density f. Then we can define the distribution of X to satisfy*

$$\mathbb{P}\{X \in A\} = p\,\mathbb{1}(0 \in A) + (1 - p)\int_A f(x)\mathrm{d}x.$$

Example 2.3.7. *Let $\Omega = [0, 1]$, and $\mathbb{P}(S) = \int_S \mathrm{d}x$, the uniform distribution on $[0, 1]$, which we denote by $\mathrm{Unif}[0, 1]$. If $\mathcal{A} = \mathbb{R}$ and $X(\omega) = \mathbb{1}(\omega \le p)$ (for some $p \in [0, 1]$), then*

$$\begin{aligned}\mathbb{P}(X = 1) &= \mathbb{P}(\{\omega \in [0, 1] \,:\, X(\omega) = 1\})\\ &= \mathbb{P}(\{\omega \in [0, 1] \,:\, \mathbb{1}(\omega \le p) = 1\})\\ &= \mathbb{P}(\{\omega \in [0, 1] \,:\, \omega \le p\})\\ &= \mathbb{P}([0, p]) = p,\end{aligned}$$

and similarly $\mathbb{P}(X = 0) = 1 - p$. So $\mathbb{P}(X \in S) = (1 - p)\mathbb{1}(0 \in S) + p\mathbb{1}(1 \in S)$ and hence $X \sim \mathrm{Bern}(p)$, a Bernoulli random variable.

Example 2.3.8. *Again consider the probability space* $([0,1], \mathbb{P})$*, where* $\mathbb{P} = \text{Unif}[0,1]$*. In the previous example we defined a single Bernoulli random variable. But we can define an uncountable family of Bernoulli random variables* $X_p \colon \Omega \to \{0,1\}$ *for each* $p \in [0,1]$ *given by* $X_p(\omega) = \mathbb{1}(\omega \leq p)$*. Notice that for* $p < q$*,* $X_p = 1 \implies X_q = 1$ *since*

$$\begin{aligned} X_p(\omega) = 1 &\iff \omega \leq p \\ &\implies \omega \leq q \\ &\iff X_q(\omega) = 1. \end{aligned}$$

Example 2.3.9. *We continue using the probability space* $([0,1], \mathbb{P})$*, where* $\mathbb{P} = \text{Unif}[0,1]$*. We can also construct two random variables* $X, Y \sim \text{Bern}(1/2)$ *such that* $\mathbb{P}\{X = Y\} = 0$*. To do, let* $X(\omega) = \mathbb{1}(\omega \leq 1/2)$ *given by* $Y(\omega) = \mathbb{1}(\omega > 1/2)$*. By definition, the probability they are equal is*

$$\mathbb{P}\{X = Y\} = \mathbb{P}(\{\omega \in \Omega \,:\, X(\omega) = Y(\omega)\}).$$

If $X(\omega) = Y(\omega) = 1$ *then we must have* $\omega \leq 1/2$ *and* $\omega > 1/2$*, which is not possible. If* $X(\omega) = Y(\omega) = 0$ *then we must have* $\omega > 1/2$ *and* $\omega \leq 1/2$*, which again is not possible. Hence,* $\{\omega \in \Omega \,:\, X(\omega) = Y(\omega)\} = \emptyset$*, so* $\mathbb{P}\{X = Y\} = \mathbb{P}(\emptyset) = 0$*.*

Exercise 2.3.2 (Identically distributed random variables that are never equal)**.** *Using the probability space* $([0,1], \mathbb{P})$*, where* $\mathbb{P} = \text{Unif}[0,1]$*, define identically distributed random variables* $X_1, \ldots, X_k$ *taking values in* $\{1, \ldots, k\}$ *such that*

(i) $\mathcal{L}_{X_i}$ *is uniform on* $\{1, \ldots, k\}$ *(that is,* $\mathbb{P}(X_i = \ell) = 1/k$ *for all* $\ell \in \{1, \ldots, k\}$*) and*

(ii) $\mathbb{P}(X_i = X_j) = 0$ *for all* $i \neq j$*.*

Example 2.3.10. *The* **Dirac measure at** x*, denoted* δ_x*, is given by* $\delta_x(S) = \mathbb{1}(x \in S)$*. In other words, if* $X \sim \delta_x$*, then* $\mathbb{P}\{X = x\} = 1$*. Using Dirac measures, we can write the Bernoulli distribution from the previous example as* $(1-p)\delta_0 + p\,\delta_1$*.*

Exercise 2.3.3 (Binomial probability measure)**.** *Similarly to Example 2.3.10, write the probability measure of a binomial random variable* $X \sim \text{Binom}(n, p)$.

Example 2.3.11 (Inversion)**.** *Consider the probability space* $([0,1], \mathbb{P})$*, where* $\mathbb{P} = \text{Unif}[0,1]$*. Say we wish to define a random variable taking values in* $\mathbb{R}$ *with* **cumulative distribution function** *c.d.f.* F *(that is, where* $F(x) = \mathbb{P}\{X \leq x\}$*). Define the* **inverse c.d.f.** *function* $F^{-1}(p) := \inf\{x : F(x) \geq p\}$*. If we define the random variable* $X \colon [0,1] \to \mathbb{R}$ *to be* $X = F^{-1}$*, then*

$$\begin{aligned}
\mathcal{L}_X((-\infty, x]) &= \mathbb{P}\{X \in (-\infty, x]\} \\
&= \mathbb{P}(\{\omega \in [0,1] \,:\, X(\omega) \in (-\infty, x]\}) \\
&= \mathbb{P}(\{\omega \in [0,1] \,:\, X(\omega) \leq x\}) \\
&= \mathbb{P}(\{\omega \in [0,1] \,:\, F^{-1}(\omega) \leq x\}) \\
&= \mathbb{P}(\{\omega \in [0, F(x)]\}) \\
&= F(x).
\end{aligned}$$

So, X *does in fact have c.d.f.* F*.*

Exercise 2.3.4 (A triangular random variable)**.** *Working with the probability space* $([0,1], \mathbb{P})$*, where* $\mathbb{P} = \text{Unif}[0,1]$*, use the approach from Example 2.3.11 to define a random variable taking values in* $[-1, 1]$ *with p.d.f.* $f(x) = 1 - |x|$*.*

Exercise 2.3.5 (A geometric random variable)**.** *Working with the probability space* $([0,1], \mathbb{P})$*, where* $\mathbb{P} = \text{Unif}[0,1]$*, use the approach from Example 2.3.11 to define a geometric random variable taking values in* $\mathbb{N}$ *with p.m.f.* $p(x) = q(1-q)^x$*, where* $q \in (0,1)$*.*

2.4 Expectation and Integration

The expectation of a discrete random variable is defined in Eq. (2.2) using its p.m.f. by *summing* over all possible values of the random variable. The expectation of a continuous random variable is defined in Eq. (2.3) using

its p.d.f. by *integrating* over all possible values. The use of summation in one and integration in the other might suggest an irreconcilable difference. But an integral is really a "continuous sum." Indeed, the symbol $\int$ is meant to resemble an elongated "S" and the so-called ***Riemann integral*** from introductory calculus is defined as the limit of a sum over increasingly small regions of the space.

Now, however, rather than a p.m.f. or p.d.f., we are using a *probability measure* to describe the distribution of a random variable. So the discrete and continuous approaches to defining the expectation no longer apply. Thus, we require a new approach to defining expectations that unifies the discrete and continuous approaches. To motivate this new definition, notice that for either a discrete or continuous random variable X defined on a set $\mathcal{A}$, for any $A \in \mathcal{P}(\mathcal{A})$,

$$\mathbb{E}\{\mathbb{1}(X \in A)\} = \mathbb{P}\{X \in A\} \tag{2.4}$$

(Check this for yourself!) Any reasonable definition of the expectation would seem to require this identity to hold, so we will require our new, more general definition of expectation to satisfy Eq. (2.4) as well.

If $X \sim \mu$, then by definition $\mathbb{P}\{X \in A\} = \mu(A)$. So, we can rewrite Eq. (2.4) as requiring that

$$\mathbb{E}\{\mathbb{1}(X \in A)\} = \mu(A). \tag{2.5}$$

But just as in the discrete and continuous cases, we want to be able to write expectations as some sort of integral (or sum) that doesn't reference the random variable itself – just the object that determines its distribution. In the discrete case the object was the p.m.f., in the continuous case the object was the p.d.f., and in this more general case the objective is the probability measure. So, we will choose to *define* our new integral to satisfy Eq. (2.5):

$$\mathbb{E}\{\mathbb{1}(X \in A)\} := \int \mathbb{1}(x \in A)\mu(\mathrm{d}x) := \mu(A). \tag{2.6}$$

We also require our new integral to have the usual properties of integrals and sums: for any constant $a \in \mathbb{R}$ and any (reasonable) functions $\phi, \psi \colon \mathcal{A} \to \mathbb{R}$,

$$\int a\,\phi(x)\mu(\mathrm{d}x) := a \int \phi(x)\mu(\mathrm{d}x) \tag{2.7}$$

and

$$\int [\phi(x) + \psi(x)]\mu(\mathrm{d}x) := \int \phi(x)\mu(\mathrm{d}x) + \int \psi(x)\mu(\mathrm{d}x). \tag{2.8}$$

The integral with these properties is called the ***Lebesgue integral***. Now generalizing Eq. (2.6), we will define the expectation of a random variable in terms of the Lebesgue integral:

$$\mathbb{E}\{\phi(X)\} := \int \phi(x)\mu(\mathrm{d}x).$$

We can also rewrite the expectation as an integral with respect to $\mathbb{P}$.

Proposition 2.4.1. *Given a probability space $(\Omega, \mathbb{P})$ and random variable $X : \Omega \to \mathcal{A}$,*

$$\mathbb{E}(X) = \int X(\omega)\mathbb{P}(\mathrm{d}\omega).$$

One heuristic way to interpret this result is that if $W \sim \mathbb{P}$, then $\mathbb{E}\{X(W)\} = \int X(\omega)\mathbb{P}(\mathrm{d}\omega) = \int x\mathcal{L}_X(\mathrm{d}x) = \mathbb{E}(X)$.

Example 2.4.2. *The Lebesgue integral with respect to the Dirac measure δ_x corresponds to evaluating the integrand at x:*

$$\int \phi(y)\delta_x(\mathrm{d}y) = \phi(x).$$

Example 2.4.3. *If X is a continuous random variable with p.d.f. f. Then its distribution is μ_f given by $\mu_f(A) = \int \mathbb{1}(x \in A)f(x)\mathrm{d}x$ and the Lebesgue integral of ϕ with respect to μ_f is equal to the Riemann integral of ϕf:*

$$\mathbb{E}\{\phi(X)\} = \int \phi(x)\mu(\mathrm{d}x) = \int \phi(x)f(x)\mathrm{d}x.$$

Exercise 2.4.1. *Use the properties of the Lebesgue integral to show that for a real-value random variable X with finite second moment and $m := \mathbb{E}(X)$,*

$$\mathrm{Var}(X) := \mathbb{E}\{(X - m)^2\} = \mathbb{E}(X^2) - m^2.$$

In addition to the more familiar integral notation, we will frequently use the more succinct notation

$$\mu(\phi) := \int \phi(x)\mu(\mathrm{d}x). \tag{2.9}$$

Which of these three notations we use will depend on context. For example, if a random variable X has already been defined, the expectation notation $\mathbb{E}\{\phi(X)\}$ can be convenient. But if we are discussing a distribution without reference to a particular random variable, then one of the integral notations that are used in Eq. (2.9) might be more convenient since they avoid the need to introduce addition symbols (namely, a random variable with the distribution of interest).

Exercise 2.4.2. *Rewrite Eqs.* (2.7) *and* (2.8) *using the notation defined in Eq.* (2.9).

2.4.1 Properties of the Lebesgue Integral

Lebesgue integrals (and hence expectations) essentially have all the same standard properties as the continuous and discrete versions (which, as we detail later, are special cases). First, note that given a constant $a \geq 0$ and probability measures μ, ν defined on $\mathcal{A}$, we can define new ***measures*** $a\mu$ and $\mu + \nu$ since we are just manipulating functions: for $A \in \mathcal{P}(\mathcal{A})$,

$$\begin{aligned}(a\mu)(A) &:= a\mu(A)\\ (\mu+\nu)(A) &:= \mu(A) + \nu(A).\end{aligned}$$

We call these *measures* rather than *probability measures* because their total mass is no longer 1. For example, $(\mu+\nu)(\mathcal{A}) = \mu(\mathcal{A}) + \nu(\mathcal{A}) = 1 + 1 = 2$. But they do satisfy condition (b) from Definition 2.3.2. It turns out that condition (b) is the only one required to define the Lebesgue integral of a measure. In particular we have

$$\begin{aligned}\int \phi(y)(a\mu)(\mathrm{d}y) &= \int a\,\phi(y)\mu(\mathrm{d}y) = a\int \phi(y)\mu(\mathrm{d}y)\\ \int \phi(y)(\mu+\nu)(\mathrm{d}y) &= \int \phi(y)\mu(\mathrm{d}y) + \int \phi(y)\nu(\mathrm{d}y).\end{aligned}$$

Or, using our more succinct notation,

$$\begin{aligned}(a\mu)(\phi) &= \mu(a\phi) = a\mu(\phi)\\ (\mu+\nu)(\phi) &= \mu(\phi) + \nu(\phi).\end{aligned}$$

Example 2.4.4. *We can now check that the Lebesgue integral definition of expectation reduces to the old definition for discrete random variables.*

For a discrete random variable X with p.m.f. $f: V \to [0,1]$, the equivalent probability measure is $\mu = \sum_{x \in V} f(x)\delta_x$. First, we can check this definition of μ is consistent with Eq. (2.1):

$$\Pr\{X \in A\} = \mu(A) = \sum_{x \in V} f(x)\delta_x(A) = \sum_{x \in V} f(x)\mathbb{1}(x \in A) = \sum_{x \in V \cap A} f(x).$$

More generally, we can check that the Lebesgue integral definition of the expectation is consistent with Eq. (2.2):

$$\begin{aligned}
\mathbb{E}\{\phi(X)\} &= \int \phi(y)\mu(\mathrm{d}y) = \int \phi(y)\left[\sum_{x \in V} f(x)\delta_x\right](\mathrm{d}y) \\
&= \sum_{x \in V} \int \phi(y)[f(x)\delta_x](\mathrm{d}y) = \sum_{x \in V} f(x) \int \phi(y)\delta_x(\mathrm{d}y) \\
&= \sum_{x \in V} f(x)\phi(x).
\end{aligned}$$

Example 2.4.5. *Let X be the random variable of biking distance from Example 2.3.6. Then, using the μ_f notation from Example 2.4.3, $\mu := \mathcal{L}_X = p\,\delta_0 + (1-p)\mu_f$, so*

$$\begin{aligned}
\mathbb{E}\{\phi(X)\} &= \int \phi(x)\mu(\mathrm{d}x) \\
&= p \int \phi(x)\delta_0(\mathrm{d}x) + (1-p) \int \phi(x)\mu_f(\mathrm{d}x) \\
&= p\,\phi(0) + (1-p) \int \phi(x)f(x)\mathrm{d}x.
\end{aligned}$$

Exercise 2.4.3. *Let $a, b, c, d \in \mathbb{R}$ be constants, let μ and ν be measures on the space $\mathcal{A}$, and let $\phi : \mathcal{A} \to \mathbb{R}$ and $\psi : \mathcal{A} \to \mathbb{R}$ be real-valued functions.*

- ***(a)*** *We can expand an expression such as $(a\mu)(b\phi + \psi)$ to the equivalent form $ab\mu(\phi) + a\mu(\psi)$. Provide a similar expansion for the expression $(a\mu + b\nu)(c\phi - d\psi)$.*
- ***(b)*** *The equivalent of the expression $(a\mu)(b\phi+\psi)$ in integral notation is $\int\{b\phi(x)+\psi(x)\}(a\mu)(\mathrm{d}x)$. Rewrite the expression $(a\mu+b\nu)(c\phi-d\psi)$*

and the expansion you provided in part ***(a)*** *in integral notation.*

Exercise 2.4.4. *Given a probability space $(\Omega, \mathbb{P})$, suppose $X : \Omega \to \mathbb{R}_+$ is a nonnegative random variable with $\mathbb{E}(X) = 1$. For each $S \in \mathcal{P}(\Omega)$, define the random variable $Y_S : \Omega \to \mathbb{R}_+$ given by $Y_S(\omega) = X(\omega)\mathbb{1}(\omega \in S)$. Define $\mathbb{Q} : \mathcal{P}(\Omega) \to \mathbb{R}_+$ by $\mathbb{Q}(S) := \mathbb{E}\{Y_S\}$.*

(a) *Show that $(\Omega, \mathbb{Q})$ is a probability space.*

(b) *Give an example showing that $\mathbb{Q}(S) = 0$ does not necessarily imply that $\mathbb{P}(S) = 0$.*

Another intuitive and useful property of the Lebesgue integral is ***monotonicity***: If $\phi(x) \leq \psi(x)$ for all $x \in \mathcal{A}$, then for any probability measure μ, the same inequality hold for the integrals of the functions with respect to μ: $\mu(\phi) \leq \mu(\psi)$.

Exercise 2.4.5. *Use monotonicity to show that for any random variable X and real-valued function ϕ, $|\mathbb{E}\{\phi(X)\}| \leq \mathbb{E}\{|\phi(X)|\}$.*

2.4.2 Multiple integrals

As in multivariate calculus, we can define multiple integrals, which turn out to be just regular Lebesgue integrals in disguise. For sets $\mathcal{A}$ and $\mathcal{B}$, define the ***Cartesian product*** $\mathcal{A} \times \mathcal{B} := \{(x, y) : x \in \mathcal{A}, y \in \mathcal{B}\}$. So, for example, $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$. Given probability spaces $(\mu, \mathcal{A})$ and $(\nu, \mathcal{B})$, for a function $\phi \colon \mathcal{A} \times \mathcal{B} \to \mathbb{R}$, we can write the multiple integrals

$$\int \left\{ \int \phi(x, y)\mu(\mathrm{d}x) \right\} \nu(\mathrm{d}y) \qquad \text{and} \qquad \int \left\{ \int \phi(x, y)\nu(\mathrm{d}y) \right\} \mu(\mathrm{d}x)$$

since the integrals in brackets are function of, respectively, y and x; hence the outer integrals are well-defined. In undergraduate multivariate calculus these two integrals are usually equal to each other, and the same is true for Lebesgue integrals. All we require is that one of the above integrals is finite when we replace $\phi(x, y)$ with its absolute value. To make this result a little more precise, and to relate the multiple integrals to a "single" Lebesgue

integral, define the ***product probability measure*** $(\mu\otimes\nu)\colon \mathcal{P}(\mathcal{A}\times\mathcal{B}) \to [0,1]$ such that for $A \in \mathcal{P}(\mathcal{A})$ and $B \in \mathcal{P}(\mathcal{B})$,

$$(\mu \otimes \nu)(A \times B) := \mu(A)\nu(B).$$

Since $\mu \otimes \nu$ is a probability measure, the integral $\int \phi(x,y)(\mu \otimes \nu)(\mathrm{d}x, \mathrm{d}y)$ is well-defined. We can now state the following key result.

Theorem 2.4.6 (Fubini–Tonelli)**.** *Given the definitions above,*

$$\begin{aligned}\int \phi(x,y)(\mu \otimes \nu)(\mathrm{d}x, \mathrm{d}y) &= \int \left\{ \int \phi(x,y)\mu(\mathrm{d}x) \right\} \nu(\mathrm{d}y) \\ &= \int \left\{ \int \phi(x,y)\nu(\mathrm{d}y) \right\} \mu(\mathrm{d}x)\end{aligned}$$

if either (a) ϕ is non-negative or (b) any of the three integrals is finite when $\phi(x,y)$ is replaced by its absolute value (e.g., if $\int |\phi(x,y)|(\mu \otimes \nu)(\mathrm{d}x, \mathrm{d}y) < \infty$).

2.5 Conditional Probabilities and Expectations

The last piece of introductory probability we need to generalize are the definitions of conditional probabilities and conditional expectations. We start by recalling what it means for two events to be independent of each other.

Definition 2.5.1. *Given a probability space $(\Omega, \mathbb{P})$, the events $S, S' \in \mathcal{P}(\Omega)$ are* **independent** *if $\mathbb{P}(S \cap S') = \mathbb{P}(S)\mathbb{P}(S')$.*

The reasoning for this definition becomes more clear when we define the conditional probability of S occurring given that S' occurs.

Definition 2.5.2. *For events $S, S' \in \mathcal{P}(\Omega)$ with $\mathbb{P}(S') > 0$, the* **conditional probability of S given S'** *is defined as $\mathbb{P}(S \mid S') := \mathbb{P}(S \cap S')/\mathbb{P}(S')$.*

Hence, if S and S' are independent, the conditional probability of S given S' is equal to the probability of S:

$$\mathbb{P}(S \mid S') = \mathbb{P}(S \cap S')/\mathbb{P}(S') = \mathbb{P}(S)\mathbb{P}(S')/\mathbb{P}(S') = \mathbb{P}(S).$$

Exercise 2.5.1 (Conditional decomposition)**.** *Show that for any $S, S' \in \mathcal{P}(\Omega)$ with $\mathbb{P}(S') \in (0,1)$, $\mathbb{P}(S) = \mathbb{P}(S \mid S')\mathbb{P}(S') + \mathbb{P}(S \mid S'^c)\mathbb{P}(S'^c)$.*

We can also define the notions of independence and conditioning for two or more random variables defined on the same probability space. We focus on the case of two random variables.

Definition 2.5.3. *The random variables $X \colon \Omega \to \mathcal{A}$ and $Y \colon \Omega \to \mathcal{B}$ are* **independent** *if for any $A \in \mathcal{P}(\mathcal{A})$ and $B \in \mathcal{P}(\mathcal{B})$, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent; that is,*

$$\mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}\{X \in A\}\mathbb{P}\{Y \in B\}.$$

The following result enables us to the define conditional distributions and conditional expectations of random variables.

Theorem 2.5.4. *Under regularity conditions, for any pair of random variables $X \colon \Omega \to \mathcal{A}$ and $Y \colon \Omega \to \mathcal{B}$, there exists a function $P_{X|Y} \colon \mathcal{B} \times \mathcal{P}(\mathcal{A}) \to [0,1]$ which has the property that for all $x \in \mathcal{A}$, the function $P_{X|Y}(x, \cdot) \colon A \mapsto P_{X|Y}(x, A)$ is a probability measure and*

$$\mathbb{P}\{X \in A, Y \in B\} = \mathbb{E}\{\mathbb{1}(Y \in B) P_{X|Y}(Y, A)\}.$$

Definition 2.5.5 (Conditional probability)**.** *For random variables $X \colon \Omega \to \mathcal{A}$ and $Y \colon \Omega \to \mathcal{B}$, the* **conditional probability of $X \in A$ given** *Y is the random variable on $(\Omega, \mathbb{P})$ given by*

$$\mathbb{P}\{X \in A \mid Y\} \colon \omega \mapsto P_{X|Y}(Y(\omega), A).$$

while the **conditional probability** *of X given $Y = y$ is the probability measure*

$$\mathbb{P}\{X \in \cdot \mid Y = y\} : A \mapsto P_{X|Y}(y, A).$$

The following result shows that the expected value of a conditional probability of the event $\{X \in S\}$ is equal to the (unconditional) probability of that event.

Proposition 2.5.6. *For any $A \in \mathcal{P}(\mathcal{A})$,*

$$\mathbb{E}[\mathbb{P}\{X \in A \mid Y\}] = \mathbb{P}\{X \in A\}.$$

Proof. It follows from Theorem 2.5.4 with $S' = \mathcal{B}$ that

$$\begin{aligned}\mathbb{E}[\mathbb{P}\{X \in S \mid Y\}] &= \mathbb{E}\{P_{X|Y}(Y, S)\} \\ &= \mathbb{E}\{\mathbb{1}(Y \in \mathcal{B}) P_{X|Y}(Y, S)\} \\ &= \mathbb{P}\{X \in S, Y \in \mathcal{B}\} \\ &= \mathbb{P}\{X \in S\}.\end{aligned}$$

□

Given random variables $X \colon \Omega \to \mathcal{A}$, $X' \colon \Omega \to \mathcal{A}'$, we have the useful identity that

$$\begin{aligned}\mathbb{P}\{X \in A, X' \in \mathcal{A}'\} &= \mathbb{P}(\{\omega \in \Omega \,:\, X \in A, X' \in \mathcal{A}\}) \\ &= \mathbb{P}(\{\omega \in \Omega \,:\, X \in A\}) \\ &= \mathbb{P}\{X \in A\}.\end{aligned}$$

The same result holds for conditional probabilities.

Proposition 2.5.7. *Given random variables $X \colon \Omega \to \mathcal{A}$, $X' \colon \Omega \to \mathcal{A}'$, and $Y \colon \Omega \to \mathcal{B}$, for all $y \in \mathcal{B}$ and $A \in \mathcal{P}(\mathcal{A})$, we have*

$$P_{X|Y}(y, A) = P_{X,X'|Y}(y, A \times \mathcal{A}').$$

Hence $\mathbb{P}\{X \in A \mid Y\} = \mathbb{P}\{X \in A, X' \in \mathcal{A}' \mid Y\}$.

Proof. We won't provide a rigorous proof, but the result essential follows from the fact that for all $B \in \mathcal{P}(\mathcal{B})$,

$$\begin{aligned}\mathbb{E}\{\mathbb{1}(Y \in B) P_{X|Y}(Y, A)\} &= \mathbb{P}\{X \in A, Y \in B\} \\ &= \mathbb{P}\{X \in A, X' \in \mathcal{A}', Y \in B\} \\ &= \mathbb{E}\{\mathbb{1}(Y \in B) P_{X|Y}(Y, A \times \mathcal{A}')\}.\end{aligned}$$

□

Example 2.5.8. *Consider random variables (X, Y) with a multivariate Gaussian joint distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{m} \in \mathbb{R}^2$ and $\boldsymbol{V} \in \mathbb{R}^{2\times 2}$. Letting $m(y) := m_1 + v_{12} v_{22}^{-1}(y - m_2)$ and $v := v_{11} - v_{12}^2 v_{22}^{-1}$, we have*

$$P_{X|Y}(y, \cdot) = \mathcal{N}(m(y), v).$$

Since $P_{X|Y}(y,\cdot)$ is the probability measure for a continuous distribution, it has a p.d.f. $p(y,\cdot)$ given by

$$p(y,x) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{\{x-m(y)\}^2}{2v}}.$$

We can use this fact to conclude that

$$\begin{aligned}\mathbb{P}\{X \in S \mid Y(\omega)\} &= P_{X|Y}(Y(\omega), S) \\ &= \int \mathbb{1}(x \in S) P_{X|Y}(Y(\omega), \mathrm{d}x) \\ &= \int \mathbb{1}(x \in S) p(Y(\omega), x) \mathrm{d}x,\end{aligned}$$

where each expression is a random variable when considered as a function of ω. For example, if $S = [a,b]$ for $a < b$, then we have

$$\mathbb{P}\{a \le X \le b \mid Y(\omega)\} = \int_a^b \frac{1}{\sqrt{2\pi v}} e^{-\frac{\{x-m(Y(\omega))\}^2}{2v}} \mathrm{d}x.$$

Exercise 2.5.2. *Given a probability space $(\Omega, \mathbb{P})$ and real-value random variables X and Y with finite second moments, show that if $m := \mathbb{E}(X)$ and $\widetilde{m} := \mathbb{E}(Y)$, then*

$$\mathrm{Cov}(X,Y) := \mathbb{E}\{(X-m)(Y-\widetilde{m})\} = \mathbb{E}(XY) - m\widetilde{m}$$

We can also use Theorem 2.5.4 to define two versions of the conditional expectation: one in which Y is unknown (and hence the conditional expectation is a random variable) and one in which Y takes on a fixed value (and hence the conditional expectation is real-valued).

Definition 2.5.9. *For random variables $X\colon \Omega \to \mathcal{A}$ and $Y\colon \Omega \to \mathcal{B}$, and a function $\phi\colon \mathcal{A} \times \mathcal{B} \to \mathbb{R}$, the* **conditional expectation of $\phi(X,Y)$ given** *Y is the random variable on $(\Omega, \mathbb{P})$ given by*

$$\mathbb{E}\{\phi(X,Y) \mid Y\}\colon \omega \mapsto \int \phi(x, Y(\omega)) P_{X|Y}(Y(\omega), \mathrm{d}x)$$

while the **conditional expectation of $\phi(X,Y)$ given** *$Y = y$ is the real value*

$$\mathbb{E}\{\phi(X,Y) \mid Y = y\} = \int \phi(x,y) P_{X|Y}(y, \mathrm{d}x).$$

Conditional expectations satisfy the so-called ***tower property***: the expected valued of the conditional expectation is equal to the (unconditional) expectation.

Proposition 2.5.10. *The conditional expectation satisfies*

$$\mathbb{E}[\mathbb{E}\{\phi(X,Y) \mid Y\}] = \mathbb{E}\{\phi(X,Y)\}$$

and, letting $\nu = \mathcal{L}_Y$*,*

$$\int \mathbb{E}\{\phi(X,Y) \mid Y = y\}\nu(\mathrm{d}y) = \mathbb{E}\{\phi(X,Y)\}.$$

The tower property also holds for multiple conditional integrals

Proposition 2.5.11. *For random variables* X, X', Y*,*

$$\mathbb{E}\big[\mathbb{E}\{\phi(X,X',Y) \mid X',Y\} \mid Y\big] = \mathbb{E}\{\phi(X,X',Y) \mid Y\}.$$

2.6 Limit Theorems

Often we are faced with a sequence of random variables $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ and we would like to know if their limiting behavior is somehow predictable.

Example 2.6.1. *Consider a sequence of i.i.d. random variables* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ *in* $\mathbb{R}^D$ *with mean* $\boldsymbol{m}$*. We would like to know if the sequence of sample averages* $\overline{\boldsymbol{X}}_k := k^{-1}\sum_{\ell=1}^k \boldsymbol{X}_k$ *converges to* $\boldsymbol{m}$*. Such a result is called a* **law of large numbers** *(LLN).*

Example 2.6.2. *Consider a sequence of i.i.d. random variables* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ *in* $\mathbb{R}^D$ *with mean* $\boldsymbol{m}$ *and covariance* $\boldsymbol{\Sigma}$*. We would like to know if the distribution of the centered and rescaled sample average* $\widetilde{\boldsymbol{X}}_k := k^{1/2}(\overline{\boldsymbol{X}}_k - \boldsymbol{m})$ *will converge. Such a result is called a* **central limit theorem** *(CLT).*

To obtain a law of large numbers or a central limit theorem, we need to define what it means to "converge." There are many possibilities. We start by considering what it means for the distributions of random variables to converge:

Definition 2.6.3 (Weak convergence)**.** *A sequence of probability measures* $\nu_1, \nu_2, \ldots$*,* **converges weakly** *to the probability measure* ν_∞ *if, for all bounded, continuous functions* ϕ*,* $\lim_{k\to\infty} \nu_k(\phi) = \nu_\infty(\phi)$*. We then write* $\nu_k \xrightarrow{w} \nu_\infty$*.*

Remark 2.6.4. *The definition of weak convergence is very general. Since it only requires the notion of a continuous function on $\mathcal{A}$ to be well-defined (that is, for $\mathcal{A}$ to have a topology), it is applicable even if $\mathcal{A}$ is not a subset of $\mathbb{R}^D$.*

We can use weak convergence to define what it means for a sequence of random variables to converge even if they are not defined on the same probability space.

Definition 2.6.5 (Convergence in distribution)**.** *A sequence of random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ is said to* **converge in distribution** *to the random variable $\boldsymbol{X}_\infty$ if $\mathcal{L}_{\boldsymbol{X}_k} \xrightarrow{w} \mathcal{L}_{\boldsymbol{X}_\infty}$. We then write $\boldsymbol{X}_k \xrightarrow{d} \boldsymbol{X}_\infty$.*

In other words, $\boldsymbol{X}_k \xrightarrow{d} \boldsymbol{X}_\infty$ if for all bounded, continuous functions ϕ, we have $\lim_{k\to\infty} \mathbb{E}\{\phi(\boldsymbol{X}_k)\} = \mathbb{E}\{\phi(\boldsymbol{X}_\infty)\}$. As suggested by the name, weak convergence (and hence convergence in distribution) serves as a minimal criteria for considering a sequence of random variables to be convergent. We can now state the classical version of the central limit theorem.

Theorem 2.6.6 (Central limit theorem for i.i.d. random variables)**.** *For any sequence of i.i.d. random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ in $\mathbb{R}^D$ with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$,*

$$k^{1/2}(\overline{\boldsymbol{X}}_k - \boldsymbol{m}) \xrightarrow{d} \boldsymbol{Y},$$

where $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

If the random variables are defined on the same probability space, there are a few additional notions of convergence which are used for laws of large numbers.

Definition 2.6.7 (Convergence in probability)**.** *A sequence of random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ on $\mathbb{R}^D$ is said to* **converge in probability** *to the random variable $\boldsymbol{X}_\infty$ if for all $\varepsilon > 0$,*

$$\lim_{k\to\infty} \mathbb{P}\{\|\boldsymbol{X}_k - \boldsymbol{X}_\infty\|_2 > \varepsilon\} = 0.$$

We then write $\boldsymbol{X}_k \xrightarrow{p} \boldsymbol{X}_\infty$.

Remark 2.6.8. *If $\boldsymbol{X}_\infty$ is a constant, then convergence in probability is well-defined even if the random variables are defined on different probability spaces.*

Definition 2.6.9 (Almost sure convergence)**.** *A sequence of random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ is said to* **converge almost surely** *(or,* **converge with probability 1***) to the random variable $\boldsymbol{X}_\infty$ if*

$$\mathbb{P}\left\{\lim_{k\to\infty} \boldsymbol{X}_k = \boldsymbol{X}_\infty\right\} = 1.$$

We then write $\boldsymbol{X}_k \overset{a.s.}{\to} \boldsymbol{X}_\infty$.

Remark 2.6.10. *Recall that $\mathbb{P}\{\lim_{k\to\infty} \boldsymbol{X}_k = \boldsymbol{X}_\infty\}$ is a shorthand for the more verbose expression*

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{k\to\infty} \boldsymbol{X}_k(\omega) = \boldsymbol{X}_\infty(\omega)\right\}\right).$$

More generally, saying a statement holds ***almost surely (a.s.)*** means the event under which the statement holds has probability 1. So, another way of writing "$\boldsymbol{X}_k \overset{a.s.}{\to} \boldsymbol{X}_\infty$" is "$\lim_{k\to\infty} \boldsymbol{X}_k = \boldsymbol{X}_\infty$ a.s." We can now state the law of large numbers.

Theorem 2.6.11 (Strong law of large numbers)**.** *For any sequence of i.i.d. random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ in $\mathbb{R}^D$ and $\boldsymbol{m} \in \mathbb{R}^D$,*

$$\overline{\boldsymbol{X}}_k \overset{a.s.}{\to} \boldsymbol{m}.$$

if and only if $\mathbb{E}(\boldsymbol{X}_1) = \boldsymbol{m}$.

Example 2.6.12. *The* **Cauchy distribution** *with location parameter y and scale parameter γ has p.d.f.*

$$p(x) = \left\{\pi\gamma\left[1 + \frac{(x-y)^2}{\gamma^2}\right]\right\}^{-1}$$

and is denoted $\text{Cauchy}(y, \gamma)$*. The mean of a Cauchy distribution is undefined and so the law of large numbers does not apply. In fact, if $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ are i.i.d.* $\text{Cauchy}(y, \gamma)$*, then $\overline{\boldsymbol{X}}_k \sim \text{Cauchy}(y, \gamma)$.*

We can relate the three types of convergence as follows:

Proposition 2.6.13. *For a sequence of random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$, the following hold:*

(i) If $\boldsymbol{X}_k \overset{a.s.}{\to} \boldsymbol{X}_\infty$, then $\boldsymbol{X}_k \overset{p}{\to} \boldsymbol{X}_\infty$.

(ii) *If* $\boldsymbol{X}_k \xrightarrow{p} \boldsymbol{X}_\infty$, *then* $\boldsymbol{X}_k \xrightarrow{d} \boldsymbol{X}_\infty$.

(iii) *If* $\boldsymbol{X}_k \xrightarrow{d} c$ *for a constant* c, *then* $\boldsymbol{X}_k \xrightarrow{p} \boldsymbol{X}_\infty$.

Example 2.6.14. *If follows from Theorem 2.6.11 and Proposition 2.6.13 that for any sequence of i.i.d. random variables* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ *in* $\mathbb{R}^D$, *if* $\mathbb{E}\boldsymbol{X}_1 = \boldsymbol{m}$, *then*

$$\overline{\boldsymbol{X}}_k \xrightarrow{p} \boldsymbol{m}.$$

This result is known as the **weak law of large numbers**, *since convergence in probability is a weaker guarantee than convergence almost surely, in the sense that almost sure convergence implies convergence in probability.*

In addition, applying a continuous function to each random variable preserves all three types of convergence:

Proposition 2.6.15 (Continuous mapping)**.** *For a sequence of random variables* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ *and continuous function* ϕ, *the following hold:*

(i) *If* $\boldsymbol{X}_k \xrightarrow{d} \boldsymbol{X}_\infty$, *then* $\phi(\boldsymbol{X}_k) \xrightarrow{d} \phi(\boldsymbol{X}_\infty)$.

(ii) *If* $\boldsymbol{X}_k \xrightarrow{p} \boldsymbol{X}_\infty$, *then* $\phi(\boldsymbol{X}_k) \xrightarrow{p} \phi(\boldsymbol{X}_\infty)$.

(iii) *If* $\boldsymbol{X}_k \xrightarrow{a.s.} \boldsymbol{X}_\infty$, *then* $\phi(\boldsymbol{X}_k) \xrightarrow{a.s.} \phi(\boldsymbol{X}_\infty)$.

2.7 Stochastic Processes

The idea of a stochastic process is easy to state.

Definition 2.7.1 (Stochastic process)**.** *A* **stochastic process** *is a collection of random variables* $\boldsymbol{X} = \{\boldsymbol{X}_t\}_{t \in \mathbb{T}}$ *[defined on a probability space* $(\Omega, \mathbb{P})$*], where each* $\boldsymbol{X}_t$ *takes values in a set* $\mathcal{A}$, *which is called the* **state space**. *The set* $\mathbb{T}$ *is called the* **index set**.

Here are a few examples of stochastic processes, many of which you have probably encountered before.

Example 2.7.2 (Random variable)**.** *Any random variable is a (rather unexciting) stochastic process. Just take* $\mathbb{T}$ *to be any singleton set such as* $\{0\}$.

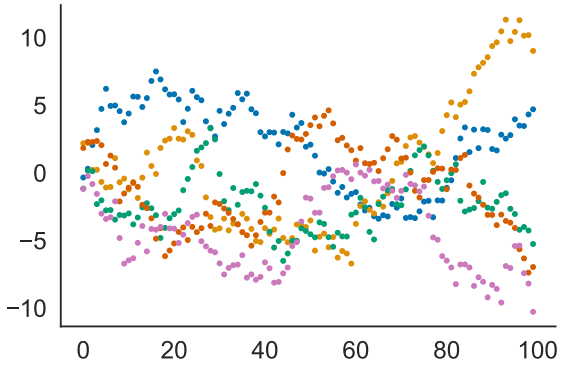


Figure 2.1: Examples of some sample paths from the random walk defined in Example 2.7.4.

Example 2.7.3 (Random vector)**.** *Let* $\mathbb{T} = \{1, \ldots, D\}$ *and* $\mathcal{A} = \mathbb{R}$*. Then* $\boldsymbol{X}$ *is a random vector in* $\mathbb{R}^D$*. For example, for any* $\mu \in \mathbb{R}^D$ *and any positive-definite matrix* $\Sigma \in \mathbb{R}^{D \times D}$*, let* $\boldsymbol{X} \sim \mathcal{N}(\mu, \Sigma)$ *be a Gaussian random vector.*

Example 2.7.4 (Random sequences and random walks)**.** *Let* $\mathbb{T} = \mathbb{N} := \{0, 1, 2, \ldots\}$ *and* $\mathcal{A} = \mathbb{R}^D$*. Then* $\boldsymbol{X}$ *is a random sequence. For example, let* $d = 1$ *and* $Z_k \overset{iid}{\sim} \mathcal{N}(0, 1)$*. Then* $\boldsymbol{X}_k = \sum_{\ell=1}^{k} Z_\ell$ *defines a random walk on* $\mathbb{R}$*. Note that* $\boldsymbol{Z} = (Z_k)_{k \in \mathbb{N}}$ *is also a stochastic process.*

Example 2.7.5 (Continuous-time processes)**.** *Stochastic processes case also take values in continuous (uncountable) index sets such as* $\mathbb{T} = \mathbb{R}_+ := [0, \infty)$*, the positive reals. We will encounter important examples of these in* **??**.

Rather than $\boldsymbol{X}_t$, sometimes we will write $\boldsymbol{X}(t)$. This alternative notation emphasizes the perspective of $\boldsymbol{X}$ as a *random function* from $\mathbb{T}$ to $\mathcal{A}$. That is, we can think of a stochastic process as an $\mathcal{A}$-valued function $\boldsymbol{X}(t, \omega)$, where $\omega \in \Omega$. If we take t as fixed, then $\boldsymbol{X}_t = \boldsymbol{X}(t, \cdot)$ is a familiar random variable again. However, if ω is fixed, then $\boldsymbol{X}(\omega) = \boldsymbol{X}(\cdot, \omega)$ is a function from $\mathbb{T}$ to $\mathcal{A}$, which is often called a *sample path*.

Example 2.7.6 (Sample paths of a random walk)**.** *Figure 2.1 shows example sample paths of the random walk defined in Example 2.7.4.*

Part II

Markov Chains

Chapter 3

Markov Chains

A Markov chain is a random sequence with the property that, given its current state, the future is independent of the past. This is formalized as the Markov property. Construction of Markov chains using an initial distribution and transition probability kernels is discussed. Stationary distributions are introduced and their possible non-uniqueness is explored. As an application, the stochastic gradient descent algorithm for optimization is introduced in detail.

Both stochastic gradient descent and Markov chain Monte Carlo involve iteratively generating a state $\boldsymbol{x}_k$, where the distribution of $\boldsymbol{x}_k$ depends only on the previous state $\boldsymbol{x}_{k-1}$. Thus, the iterates form a type of stochastic process called a ***Markov chain***. This chapter is dedicated to formally defining Markov chains and describing some basic properties.

3.1 What is a Markov Chain?

Consider a discrete-time stochastic process $\boldsymbol{X} = \{\boldsymbol{X}_k\}_{k\in\mathbb{N}}$ on state space $\mathcal{A}$. In applications we will usually assume that $\mathcal{A} = \mathbb{R}^D$. But the following discussions will mostly not depend on that assumption.

Definition 3.1.1 (Markov chain)**.** *The process $\boldsymbol{X}$ is a* **Markov chain** *if it satisfies the* **Markov condition***: for all $k \in \mathbb{N}$, all $A \subseteq \mathcal{A}$,and all $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_k \in \mathcal{A}$,*

$$\mathbb{P}\{\boldsymbol{X}_{k+1} \in A \mid \boldsymbol{X}_0 = \boldsymbol{x}_0, \ldots, \boldsymbol{X}_k = \boldsymbol{x}_k\} = \mathbb{P}\{\boldsymbol{X}_{k+1} \in A \mid \boldsymbol{X}_k = \boldsymbol{x}_k\}. \tag{3.1}$$

In other words, a Markov chain satisfies the property that if you know its state at some time k, its future states (after k) are independent of the

past (before k). In fact, the Markov condition is equivalent to two other conditions:

$$\begin{aligned}&\text{for all } k \in \mathbb{N}, \text{ all } A \subseteq \mathcal{A}, \text{ all } \boldsymbol{x}_0, \boldsymbol{x}_1, \ldots \in \mathcal{A}, \text{ and all}\\ &0 \leq k_1 < \cdots < k_i \leq k,\\ &\mathbb{P}\{\boldsymbol{X}_{k+1} \in A \mid \boldsymbol{X}_{k_1} = \boldsymbol{x}_{k_1}, \ldots, \boldsymbol{X}_{k_i} = \boldsymbol{x}_{k_i}\} = \mathbb{P}\{\boldsymbol{X}_{k+1} \in A \mid \boldsymbol{X}_{k_i} = \boldsymbol{x}_{k_i}\}\end{aligned} \tag{3.2}$$

and

$$\begin{aligned}&\text{for all } k, m \in \mathbb{N}, \text{ all } A \subseteq \mathcal{A}, \text{ and all } \boldsymbol{x}_0, \boldsymbol{x}_1, \ldots \in \mathcal{A},\\ &\mathbb{P}\{\boldsymbol{X}_{k+m} \in A \mid \boldsymbol{X}_0 = \boldsymbol{x}_0, \ldots, \boldsymbol{X}_k = \boldsymbol{x}_k\} = \mathbb{P}\{\boldsymbol{X}_{k+m} \in A \mid \boldsymbol{X}_k = \boldsymbol{x}_k\}.\end{aligned} \tag{3.3}$$

Exercise 3.1.1 (The Markov condition)**.** *Show that* ***(a)*** *Eq.* (3.2) $\implies$ *Eq.* (3.3) *and* ***(b)*** *Eq.* (3.3) $\implies$ *Eq.* (3.1)*.*

The Markov condition implies that the evolution of a Markov chain can be fully described by the ***transition probabilities***

$$\mathbb{P}\{\boldsymbol{X}_{k+1} \in A \mid \boldsymbol{X}_k = \boldsymbol{x}_k\}.$$

In general these transition probabilities depend on k. However, often there is no such dependence.

Definition 3.1.2 (Homogeneity)**.** *A Markov chain $\boldsymbol{X}$ is* **homogenous** *if for all $A \subseteq \mathcal{A}$ and $\boldsymbol{x} \in \mathcal{A}$, its transition probabilities satisfy*

$$\mathbb{P}\{\boldsymbol{X}_{k+1} \in A \mid \boldsymbol{X}_k = \boldsymbol{x}\} = \mathbb{P}\{\boldsymbol{X}_1 \in A \mid \boldsymbol{X}_0 = \boldsymbol{x}\}.$$

Otherwise the chain is called **inhomogenous**.

Example 3.1.3 (A random sequence)**.** *Any sequence of independent random variables $\boldsymbol{X}$ is a Markov chain since then each random variable is independent of previous ones: $\mathbb{P}\{\boldsymbol{X}_{k+1} \in \cdot \mid \boldsymbol{X}_k = \boldsymbol{x}_k, \ldots, \boldsymbol{X}_0 = \boldsymbol{x}_0\} = \mathcal{L}_{\boldsymbol{X}_{k+1}}$. The sequence is homogenous if the random variables are identically distributed. For example, if $\boldsymbol{X}_k \sim \mathcal{N}(k, 1)$ independent for each $k \in \mathbb{N}$, then the chain is inhomogenous. On the other hand, if for constants m and v we have $\boldsymbol{X}_k \sim \mathcal{N}(m, v)$ independent for each $k \in \mathbb{N}$, then the chain is homogenous.*

Example 3.1.4 (A random walk)**.** *The random walk $\boldsymbol{X}$ from Example 2.7.4 is a homogenous Markov process with $\mathbb{P}\{\boldsymbol{X}_{k+1} \in \cdot \mid \boldsymbol{X}_k = \boldsymbol{x}\} = \mathcal{N}(\boldsymbol{x}, 1)$; for, equivalently, we can write this conditional distribution as*

$$\boldsymbol{X}_{k+1} \mid \boldsymbol{X}_k = \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}, 1).$$

Example 3.1.5 (AR(1) process)**.** *Taking $\mathcal{A} = \mathbb{R}^D$, the* **first-order autoregressive [AR(1)] process** *is given by*

$$\boldsymbol{X}_k = \alpha \boldsymbol{X}_{k-1} + \varepsilon_k,$$

where $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_k)_{k\in\mathbb{N}}$ is an i.i.d. sequence of random variables satisfying $\mathbb{E}(\boldsymbol{\varepsilon}_0) = 0$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}_0) = \boldsymbol{\Sigma}_\varepsilon$. Letting $A - \boldsymbol{y} := \{\boldsymbol{x} - \boldsymbol{y} \mid \boldsymbol{x} \in A\}$, we can see it is a homogenous Markov process with

$$\begin{aligned}\mathbb{P}\{\boldsymbol{X}_k \in A \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}\} &= \mathbb{P}\{\alpha \boldsymbol{X}_{k-1} + \varepsilon_k \in A \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}\} \\ &= \mathbb{P}\{\varepsilon_k \in A - \alpha\boldsymbol{x}\}.\end{aligned}$$

For example, if $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\varepsilon)$, then

$$\boldsymbol{X}_k \mid \boldsymbol{X}_{k-1} = \boldsymbol{x} \sim \mathcal{N}(\alpha\boldsymbol{x}, \boldsymbol{\Sigma}_\varepsilon).$$

3.2 Probability Kernels

Recall from Section 2.5 that, following Definition 2.5.5, the conditional distribution $\mathbb{P}\{\boldsymbol{X}_k \in A \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}\}$ is defined in terms of the function $P_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}} \colon \mathcal{A} \times \mathcal{P}(\mathcal{A}) \to [0,1]$. It follows from Theorem 2.5.4 that keeping $\boldsymbol{x} \in \mathcal{A}$ fixed, $P_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}(\boldsymbol{x}, \cdot)$ is a probability measure. In the context of Markov chains, we will find it much more convenient to work with $P_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}$, which we call a *probability kernel*:

Definition 3.2.1 (Probability kernel)**.** *A* **probability kernel** *(also called a* **Markov kernel***) is a function $P \colon \mathcal{A} \times \mathcal{P}(\mathcal{A}) \to [0,1]$ which has the property that, for all $\boldsymbol{x} \in \mathcal{A}$, the function $P(\boldsymbol{x}, \cdot) \colon A \mapsto P(\boldsymbol{x}, A)$ is a probability measure.*

Given a Markov chain, we can define the ***kth transition kernel*** P_k given by

$$P_k(\boldsymbol{x}, A) := P_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}(\boldsymbol{x}, A) = \mathbb{P}(\boldsymbol{X}_k \in A \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}).$$

The distribution of the Markov chain is fully described by the initial distribution $\nu_0 := \mathcal{L}_{\boldsymbol{X}_0}$ and the transition kernels $(P_k)_{k\geq 1}$.

Example 3.2.2 (Finite-state Markov chains)**.** *Introductory stochastic processes courses (and some probability theory courses) cover discrete-state Markov chains, where $\mathcal{A} \subseteq \mathbb{Z}$. For concreteness, consider the case of*

$\mathcal{A} = [D] = \{1, \dots, D\}$. We can write the distribution of the Markov chain at index k as a vector $\boldsymbol{\pi}_k \in [0,1]^D$, where

$$\pi_{k,d} := \Pr\{\boldsymbol{X}_k = d\}.$$

Note that the sum of the components of $\boldsymbol{\pi}_k$ must be 1:

$$\sum_{d \in [D]} \pi_{k,d} = \sum_{d \in [D]} \Pr\{\boldsymbol{X}_k = d\} = \Pr\{\boldsymbol{X}_k \in [D]\} = 1$$

since $\mathcal{A} = [D]$. We can also express distribution of $\boldsymbol{X}_k$ using $\boldsymbol{\pi}_k$:

$$\mathcal{L}_{\boldsymbol{X}_k} = \sum_{d \in [D]} \pi_{k,d} \delta_d.$$

We can summarize the transition kernel P_k using a **transition matrix** *$\boldsymbol{K}_k \in [0,1]^{D \times D}$, where*

$$K_{k,d,d'} := \mathbb{P}\{\boldsymbol{X}_k = d' \mid \boldsymbol{X}_{k-1} = d\} = P_k(d, \{d'\}).$$

Each row of the transition matrix must sum to 1 by the properties of the transition kernel and probability measures:

$$\sum_{d' \in [D]} K_{k,d,d'} = \sum_{d' \in [D]} P_k(d, \{d'\}) = P_k(d, [D]) = 1.$$

We can also express the transition kernel in terms of the transition matrix:

$$P_k(d, \cdot) = \sum_{d' \in [D]} K_{k,d,d'} \delta_{d'}.$$

Exercise 3.2.1 (SGD transition kernel)**.** *Give the transition kernel for the SGD update from Eq. (1.4) when* ***(a)*** *$B = 1$,* ***(b)*** *$B = 2$, and* ***(c)*** *B is any positive integer.*

3.2.1 Conditional distributions

We would like to be able to write arbitrary conditional distributions of the form $\mathbb{P}\{\boldsymbol{X}_k \in A \mid \boldsymbol{X}_\ell = \boldsymbol{x}\}$ for $0 \le \ell < k$. Before writing out the general case, we build some intuitions by considering a finite-state Markov chain.

Example 3.2.3 (Conditional distributions of a finite-state Markov chain)**.** *Following the set-up and notation from Example 3.2.2, to find the two-step conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_0$, we can multiply the corresponding matrices together: letting*

$$\boldsymbol{K}_{0\to 2} := \boldsymbol{K}_1 \boldsymbol{K}_2,$$

we have

$$\begin{aligned}
&K_{0\to 2, d_0, d_2} \\
&= \sum_{d_1 \in [D]} K_{1,d_0,d_1} K_{2,d_1,d_2} \quad \textit{definition of matrix multiplication} \\
&= \sum_{d_1 \in [D]} \mathbb{P}\{\boldsymbol{X}_1 = d_1 \mid \boldsymbol{X}_0 = d_0\} \mathbb{P}\{\boldsymbol{X}_2 = d_2 \mid \boldsymbol{X}_1 = d_1\} \quad \textit{definitions of } \boldsymbol{K}_1 \textit{ and } \boldsymbol{K}_2 \\
&= \sum_{d_1 \in [D]} \mathbb{P}\{\boldsymbol{X}_1 = d_1 \mid \boldsymbol{X}_0 = d_0\} \mathbb{P}\{\boldsymbol{X}_2 = d_2 \mid \boldsymbol{X}_1 = d_1, \boldsymbol{X}_0 = d_0\} \quad \textit{Markov assumption} \\
&= \sum_{d_1 \in [D]} \mathbb{P}\{\boldsymbol{X}_2 = d_2, \boldsymbol{X}_1 = d_1 \mid \boldsymbol{X}_0 = d_0\} \\
&= \mathbb{P}\{\boldsymbol{X}_2 = d_2 \mid \boldsymbol{X}_0 = d_0\}
\end{aligned}$$

In other words, we can find the conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_0$ by multiplying the transition matrices for times 0 to 1 and 1 to 2 together. Using the same idea, we have the more general relationship that if

$$\boldsymbol{K}_{\ell\to k} := \boldsymbol{K}_{\ell+1} \boldsymbol{K}_{\ell+2} \cdots \boldsymbol{K}_k,$$

then

$$K_{\ell\to k, d, d'} = \mathbb{P}\{\boldsymbol{X}_k = d' \mid \boldsymbol{X}_\ell = d_d\}.$$

In the general case, the conditional distribution $\mathbb{P}(\boldsymbol{X}_2 \in A \mid \boldsymbol{X}_0 = \boldsymbol{x})$ takes similar form, which integration replacing matrix multiplication (we omit the derivation):

$$\int P_2(\boldsymbol{x}_1, A_2) P_1(\boldsymbol{x}_0, \mathrm{d}\boldsymbol{x}_1) = P_{\boldsymbol{X}_2 | \boldsymbol{X}_0}(\boldsymbol{x}_0, A_2).$$

We call the right-hand side the ***composition*** of the 1st and 2nd transition kernels.

Definition 3.2.4 (Composition of probability kernels)**.** *Given kernels P and P', the* **kernel composition** *PP' is the probability kernel such that for $\boldsymbol{x} \in \mathcal{A}$ and $A \subseteq \mathcal{A}$,*

$$(PP')(\boldsymbol{x}, A) := \int P'(\boldsymbol{y}, A) P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}).$$

Lemma 3.2.5. *The composition PP' is a probability kernel.*

Proof. Taking $\boldsymbol{x} \in \mathcal{A}$ as fixed, we check the requirements of Definition 2.3.2 are satisfied by $PP'(\boldsymbol{x}, \cdot)$. For pairwise disjoint sets $A_1, A_2, \cdots \subseteq \mathcal{A}$,

$$\begin{aligned}
(PP')(\boldsymbol{x}, \cup_{i\geq 1} A_i) &= \int P'(\boldsymbol{y}, \cup_{i\geq 1} A_i) P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \quad \text{by definition of } PP' \\
&= \int \sum_{i=1}^{\infty} P'(\boldsymbol{y}, A_i) P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \quad \text{because } P' \text{ is a probability kernel} \\
&= \sum_{i=1}^{\infty} \int P'(\boldsymbol{y}, A_i) P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \quad \text{by monotone convergence theorem} \\
&= \sum_{i=1}^{\infty} (PP')(\boldsymbol{x}, A_i) \quad \text{by definition of } PP'.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
(PP')(\boldsymbol{x}, \mathcal{A}) &= \int P'(\boldsymbol{y}, \mathcal{A}) P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \\
&= \int 1 P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \quad \text{because } P' \text{ is a probability kernel} \\
&= 1 \quad \text{because } P \text{ is a probability kernel.}
\end{aligned}$$

□

We can use composition repeatedly. For example, $((P_1P_2)P_3)(\boldsymbol{x}, A) = \Pr(\boldsymbol{X}_3 \in A \mid \boldsymbol{X}_0 = \boldsymbol{x})$. However, the notation is again getting clumsy due to all the parentheses. Luckily, composition is associative, so we can write $P_1P_2P_3(\boldsymbol{x}, A)$ or even $P_1P_2P_3P_3(\boldsymbol{x}, A)$ without any ambiguity.

Lemma 3.2.6 (Associativity of kernel compositions)**.** *For any probability kernels P_1, P_2, and P_3, the equality $(P_1P_2)P_3 = P_1(P_2P_3)$ holds.*

Proof. Using the definition of composition repeatedly and using Fubini's theorem, for any $\boldsymbol{x} \in \mathcal{A}$ and $A \subseteq \mathcal{A}$, we have

$$\begin{aligned}((P_1P_2)P_3)(\boldsymbol{x}, A) &= \int P_3(\boldsymbol{y}, A)(P_1P_2)(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \\ &= \int P_3(\boldsymbol{y}, A) \int P_2(\boldsymbol{y}', \mathrm{d}\boldsymbol{y}) P_1(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}') \\ &= \int P_1(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}') \int P_3(\boldsymbol{y}, A) P_2(\boldsymbol{y}', \mathrm{d}\boldsymbol{y}) \\ &= \int P_1(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}')(P_2P_3)(\boldsymbol{y}', A) \\ &= (P_1(P_2P_3))(\boldsymbol{x}, A).\end{aligned}$$

□

We have now shown that kernel composition behaves very similarly to matrix multiplication. In fact, as hinted at above, we can view our results about kernel compositions as generalizing our derivations in Example 3.2.3 for finite state spaces using matrix operations. Returning to our original goal on writing arbitrary conditional distributions, we can now conclude that

$$\Pr\{\boldsymbol{X}_k \in A_k \mid \boldsymbol{X}_\ell = \boldsymbol{x}_\ell\} = P_{\ell+1}P_{\ell+2} \cdots P_{k-1}P_k(\boldsymbol{x}_\ell, A_k).$$

For a homogenous Markov chain, since $P_k = P_1$ for all $k \geq 1$, rather than writing the composition of $P := P_1$ multiple times as PP, PPP, etc., we can instead take inspiration from matrix multiplication and recursively define the compact notation $P^\ell := P^{\ell-1}P$ with the base case $P^1 := P$.

3.2.2 Marginal distributions

We would also like to be able to compute the marginal distributions $\nu_k := \mathcal{L}_{\boldsymbol{X}_k}$ of the Markov chain iterates. Again starting with the simplest case, if the marginal distribution ν_0 is known, then using definitions and Proposition 2.5.6, we have

$$\nu_1(A) = \Pr\{\boldsymbol{X}_1 \in A\} = \mathbb{E}[\Pr\{\boldsymbol{X}_1 \in A \mid \boldsymbol{X}_0\}] = \int P_1(\boldsymbol{x}_0, A)\nu_0(\mathrm{d}\boldsymbol{x}_0).$$

We refer to integral as as the ***composition of*** ν_0 ***and*** P_1.

Definition 3.2.7 (Composition of a distribution and probability kernel)**.** *Given probability measure μ and probability kernel P, the* **distribution–kernel composition** *μP is the measure such that for $A \subseteq \mathcal{A}$,*

$$(\mu P)(A) := \int P(\boldsymbol{y}, A)\mu(\mathrm{d}\boldsymbol{y}).$$

Lemma 3.2.8. *The composition μP is a probability measure.*

Exercise 3.2.2. *Prove Lemma 3.2.8.*

We can combine a distribution–kernel composition with kernel compositions to obtain arbitrary marginals. For example, $\mathcal{L}_{\boldsymbol{X}_2} = (\nu_0 P_1)P_2$. Or, more generally, $\mathcal{L}_{\boldsymbol{X}_{k+2}} = (\nu_k P_{k+1})P_{k+2}$. Since, as stated in the next lemma, the two compositions are associative, we can drop the parentheses and write $\mathcal{L}_{\boldsymbol{X}_{k+2}} = \nu_k P_{k+1} P_{k+2}$ without any ambiguity.

Lemma 3.2.9 (Associativity of distribution and kernel compositions)**.** *Show that $(\mu P_1)P_2 = \mu(P_1 P_2)$.*

Exercise 3.2.3. *Prove Lemma 3.2.9.*

Exercise 3.2.4. *For a homogenous Markov chain with transition kernel P, for $k > \ell \geq 0$, use compositions to write ν_k in terms of P and ν_ℓ.*

3.2.3 Expectations

The final operation we would like to be able to do using kernels is compute expectations. By definition,

$$\mathbb{E}\{\phi(\boldsymbol{X}_1) \mid \boldsymbol{X}_0 = \boldsymbol{x}_0\} = \int \phi(\boldsymbol{y}) P_1(\boldsymbol{x}_0, \mathrm{d}\boldsymbol{y}). \tag{3.4}$$

We have been using the notation $\mu(\phi)$ to denote $\mathbb{E}\{\phi(\boldsymbol{X})\}$ when $\boldsymbol{X} \sim \mu$. We will use the similar notation $P_1\phi(\boldsymbol{x}_0)$ to denote the right-hand side of Eq. (3.4).

Definition 3.2.10 (Expectations with probability kernels)**.** *Given a probability kernel P and function ϕ, the* **conditional kernel expectation** *$P\phi$ is the function such that for $\boldsymbol{x} \in \mathcal{A}$,*

$$(P\phi)(\boldsymbol{x}) := \int \phi(\boldsymbol{y}) P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}).$$

When combined with compositions there is no ambiguity if we do not write parentheses. So, for example, for $\phi : \mathcal{A} \to \mathbb{R}$,

$$\begin{aligned} \nu_0 P_1 P_2 \phi &= \nu_0 P_1 P_2(\phi) \\ &= \nu_0(P_1 P_2 \phi) \\ &= \int \mathbb{E}\{\phi(\boldsymbol{X}_2) \mid \boldsymbol{X}_0 = \boldsymbol{x}_0\} \nu_0(\mathrm{d}\boldsymbol{x}_0) \\ &= \mathbb{E}[\mathbb{E}\{\phi(\boldsymbol{X}_2) \mid \boldsymbol{X}_0 = \boldsymbol{x}_0\}] \\ &= \mathbb{E}\{\phi(\boldsymbol{X}_2)\} \end{aligned} \tag{3.5}$$

is the expected value of $\phi(\boldsymbol{X}_2)$.

Exercise 3.2.5. *For a homogenous Markov chain with transition kernel P and initial distribution ν_0, write each of the following as a probability, expectation, or conditional expectation, similarly to Eq. (3.5): (a) $\nu_0 P^k$, (b) $P^k\phi$, and (c) $\nu_0 P^k \phi$.*

3.3 Stationary Distributions

In stochastic optimization it can be useful to characterize the limiting distribution of the iterates when using constant step size. Or, for Markov chain Monte Carlo, we want to construct transition kernels that will lead to the iterates having a desired distribution π. To answer these types of questions, it is necessary to determine for which distribution (or distributions) a Markov chain will remain constant.

Definition 3.3.1 (Invariant / stationary distributions)**.** *The probability measure ν is an* **invariant distribution** *of the homogenous Markov chain if $\nu = \nu P$; that is, when the distribution of $\boldsymbol{X}_k$ is ν, the distribution of $\boldsymbol{X}_{k+1}$ is also ν. The invariant distribution is also called the* **stationary distribution** *and if the distribution of $\boldsymbol{X}_0$ is ν, the Markov chain is said to be* **at stationarity**.

Example 3.3.2 (Stationary distribution of a finite-state Markov chain)**.** *For a finite-state Markov chain with transition matrix $\boldsymbol{K} \in [0,1]^{D \times D}$, the invariant distribution condition is $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{K}$ for $\boldsymbol{\pi} \in [0,1]^D$ satisfying $\sum_{d=1}^D \pi_d = 1$. Thus, the invariant distributions are exactly the left eigenvectors of $\boldsymbol{K}$ with eigenvalue 1. For example, if $\boldsymbol{K} = \boldsymbol{I}$ (the identity matrix), the probability of a staying in the same state is always 1. Also, $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{I} = \boldsymbol{\pi}K$ for all $\boldsymbol{\pi}$, so every distribution is an invariant distribution of a Markov chain with this transition matrix! On the other hand, if $K_{d,d'} = 1/D$ for all $d, d' \in [D]$, its only eigenvector with eigenvalue 1 is the constant vector $\pi_d = 1/D$. Since eigenvectors are always determined up to a multiplicative constant, we choose to scale them such that they represent valid probability distributions (i.e., the sum of the components equals 1).*

Example 3.3.3 (AR(1) process, continued)**.** *In the setting of Example 3.1.5, take $\boldsymbol{\varepsilon}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon)$. If $\alpha = 0$ then $\boldsymbol{X}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon)$. So, assume $\alpha \neq 0$. Then if $\mu = \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$ and $\boldsymbol{X}' \sim \mu$, it follows that $\alpha\,\boldsymbol{X}' \sim \mathcal{N}(\alpha\boldsymbol{m}, \alpha^2\boldsymbol{S})$ and hence $X \sim \mu P$ has distribution $\mathcal{N}(\alpha\,\boldsymbol{m}, \alpha^2\,\boldsymbol{S} + \boldsymbol{\Sigma}_\varepsilon)$. By induction, it follows that*

$$\mu P^k = \mathcal{N}\left(\alpha^k \boldsymbol{m}, \alpha^{2k}\boldsymbol{S} + \sum_{\ell=0}^{k-1} \alpha^{2\ell}\boldsymbol{\Sigma}_\varepsilon\right). \tag{3.6}$$

We can find an invariant distribution by setting $\mu = \mu P$, which yields the requirement that

$$\mathcal{N}(\boldsymbol{m}, \boldsymbol{S}) = \mathcal{N}(\alpha\,\boldsymbol{m}, \alpha^2\,\boldsymbol{S} + \boldsymbol{\Sigma}_\varepsilon).$$

Hence $\boldsymbol{m} = \alpha\boldsymbol{m}$*, which implies* $\boldsymbol{m} = \mathbf{0}$ *since* $\alpha \neq 0$*, and* $\boldsymbol{S} = \alpha^2\boldsymbol{S} + \Sigma_\varepsilon$*, which implies* $\boldsymbol{S} = (1-\alpha^2)^{-1}\Sigma_\varepsilon$ *as long as* $|\alpha| < 1$*. Hence, if* $\alpha \in (-1, 1)$*, the Gaussian AR(1) process has invariant distribution* $\mathcal{N}(\mathbf{0}, \{1-\alpha^2\}^{-1}\boldsymbol{\Sigma}_\varepsilon)$*.*

Exercise 3.3.1. *Prove Eq. (3.6) using induction.*

Exercise 3.3.2. *In the setting of Example 3.3.3, prove that the distribution of the AR(1) process converges to the stationary distribution* $\nu = \mathcal{N}(\mathbf{0}, \{1-\alpha^2\}^{-1}\boldsymbol{\Sigma}_\varepsilon)$*. In particular, show that for all* $A \subseteq \mathbb{R}^D$*,* $\lim_{k\to\infty} \mu P^k(A) = \nu(A)$*.* [Hint: Proving convergence in distribution for Gaussian distributions is equivalent to showing convergence of the mean and covariance.]

Exercise 3.3.3 (Non-uniqueness of invariant distributions)**.** *Show that* ***any*** *probability distribution* ν *is an invariant distribution of the Markov chain with transition kernel given by* $P(\boldsymbol{x}, A) = \delta_{\boldsymbol{x}}(A)$*.*

Chapter 4

Convex Analysis and Taylor Approximation

With Applications to Error Analysis of SGD

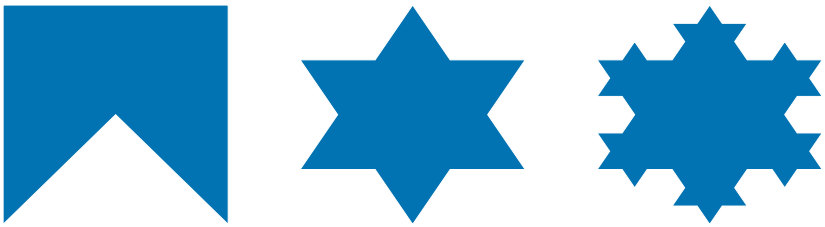
Introduction to two heavily used tools in the analysis of optimization and sampling algorithms: convex analysis and bounds on the error of Taylor series approximations. Key concepts include (strong) convexity, strong smoothness, and co-coercivity. Results are given about operations on and compositions of convex functions, and implications of (strong) convexity and strong smoothness. As an application, develops error analyses of stochastic gradient descent with constant step size (last-iterate and iterate averaging) and decreasing step size (last-iterate).

4.1 Convex Sets and Functions

When analyzing stochastic optimization and sampling algorithms, we must impose ***regularity conditions***. For stochastic optimization these conditions concern the function being minimized and the stochastic gradients being used. For sampling the conditions concern the target distribution. The theory of ***convex analysis*** provides a quite general – and very fruitful – approach to defining such regularity conditions that hold in real-world problems. Combined with very classical results on the error of Taylor series function approximations (Section 4.4), we will see that we can obtain quantitative, finite-iteration error bounds for stochastic gradient descent (SGD) with and without iterate averaging.

Our starting point, however, concerns sets rather than functions.

(a) Three convex sets.



(b) Three non-convex sets.

Figure 4.1: Examples of convex and non-convex sets.

Definition 4.1.1 (Convex sets)**.** *A set $A \subseteq \mathbb{R}^D$ is* **convex** *if for all $\boldsymbol{x}, \boldsymbol{y} \in A$ and $t \in [0,1]$, $t\boldsymbol{x} + (1-t)\boldsymbol{y} \in A$.*

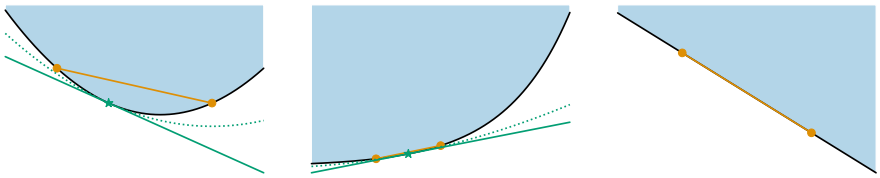
Figure 4.1 gives examples of sets in $\mathbb{R}^2$ that are convex and non-convex.

Definition 4.1.2 (Convex and concave functions)**.** *Given a convex set $\mathcal{A} \subseteq \mathbb{R}^D$, a function $\phi\colon \mathcal{A} \to \mathbb{R} \cup \{-\infty, +\infty\}$ is* **convex** *if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$ and $t \in [0,1]$,*

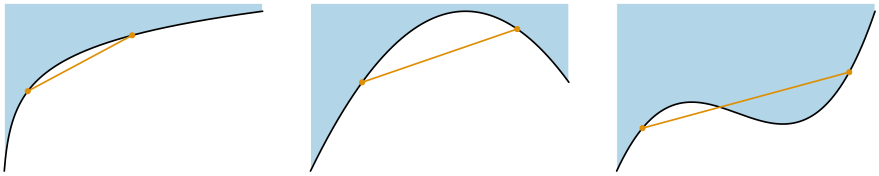
$$\phi(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq t\phi(\boldsymbol{x}) + (1-t)\phi(\boldsymbol{y}). \tag{4.1}$$

If in addition if Eq. (4.1) holds without equality (i.e., with $a <$ rather than $a \leq$) when $t \notin \{0,1\}$ and $\boldsymbol{x} \neq \boldsymbol{y}$, then ϕ is **strictly convex**. *If $-\phi$ is convex, then we say that ϕ is* **concave**.

The righthand side of Eq. (4.1) represents the line L connecting $(\boldsymbol{x}, \phi(\boldsymbol{x}))$ to $(\boldsymbol{y}, \phi(\boldsymbol{y}))$. Thus, a function is convex if and only if for each point $\boldsymbol{y}$ on the line connecting $\boldsymbol{x}$ and $\boldsymbol{y}$, the function value $f(\boldsymbol{y})$ is less than or equal to corresponding value of L at $\boldsymbol{y}$. This requirement is illustrated in Fig. 4.2, along with some additional graphical examples of functions that are convex,



(a) Three convex functions. The third function (a line) is also concave.



(b) Three non-convex functions. The first two functions are concave. The last function neither concave nor convex.

Figure 4.2: Examples of convex and non-convex functions $\phi(\boldsymbol{x})$. For the convex functions, the blue shaded region above the function is a convex set. For concave functions, the white region below the function is a convex set. The orange lines show the linear upper bound on the right-hand side of Eq. (4.1) holds for convex functions but not non-convex functions. The orange dots denote the points $(\boldsymbol{x}, \phi(\boldsymbol{x}))$ and $(\boldsymbol{y}, \phi(\boldsymbol{y}))$. The green solid line illustrates the linear lower bound in Eq. (4.2) (with the green star denoting the point $(\boldsymbol{y}, \phi(\boldsymbol{y}))$). The linear bound holds for all convex functions. The green dotted line illustrates the quadratic lower bound in Eq. (4.3).

concave, or neither. Notably, the only functions that are concave and convex are lines.

Exercise 4.1.1. *Show that a function ϕ is convex and concave if and only if $\phi(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} + b$ for some $\boldsymbol{a} \in \mathbb{R}^D$ and $b \in \mathbb{R}$.*

Although we will not make use of it going forward, the following result makes the connection between convex sets and functions explicit, as illustrated in Fig. 4.2.

Proposition 4.1.3. *A function ϕ is convex if and only its* epigraph $\{(\boldsymbol{x}, a) \in \mathcal{A} \times \mathbb{R} : a \geq \phi(\boldsymbol{x})\}$ *is a convex set.*

Example 4.1.4 (Convex functions)**.** *Functions that are convex on $\mathcal{A} = \mathbb{R}$:*

1. *The linear function* $\phi(x) = x$
2. *The quadratic function* $\phi(x) = x^2$
3. *The quartic function* $\phi(x) = x^4$
4. *The exponential function* $\phi(x) = \exp(x)$

Functions that are convex on $\mathcal{A} = [0, \infty)$*:*

1. *The negative logarithm* $\phi(x) = -\log(x)$
2. *The cubic function* $\phi(x) = x^3$
3. *The inverse function* $\phi(x) = x^{-1}$

Functions that are convex on $\mathcal{A} = \mathbb{R}^D$*:*

1. *For any vector* $\boldsymbol{b} \in \mathbb{R}^D$*, the linear function* $\phi(\boldsymbol{x}) = \boldsymbol{b}^\top \boldsymbol{x}$
2. *For any positive semidefinite matrix* $\boldsymbol{A} \in \mathbb{R}^{D \times D}$*, the quadratic form* $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$*.*
3. *For any* $p \geq 1$*, the p-norm function* $\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|_p$*.*

Since we will deal exclusively with differentiable functions, it will often be useful to leverage an equivalent characterization of convexity. For vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^D$, we will sometimes use the ***inner product*** notation $\langle \boldsymbol{a}, \boldsymbol{b} \rangle := \boldsymbol{a}^\top \boldsymbol{b} = \boldsymbol{a} \cdot \boldsymbol{b}$ to denote the dot product.

Proposition 4.1.5. *A differentiable function* $\phi \colon \mathcal{A} \to \mathbb{R}$ *is convex if and only for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$*,*

$$\phi(\boldsymbol{x}) \geq \phi(\boldsymbol{y}) + \langle \phi'(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle. \tag{4.2}$$

The right-hand side of Eq. (4.2) can be viewed as the linear approximation to $\phi(\boldsymbol{x})$ using the first-order Taylor series at $\boldsymbol{y}$, as illustrated in Fig. 4.2. Thus, we have the useful analytical fact that the linear approximation of a convex function provides a lower bound on the function's value.

Exercise 4.1.2 (Checking that functions are convex)**.** *Verify using any of the equivalent definitions that the functions listed in Example 4.1.4 are convex.*

A more stringent requirement than convexity is that of ***strong convexity***, which guarantees a *quadratic* lower bound.

Definition 4.1.6 (Strong convexity)**.** *For $\mu > 0$, a function $\phi\colon \mathcal{A} \to \mathbb{R}$ is* **μ-strongly convex** *if any of the following equivalent conditions holds:*

1. *For all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$ and $t \in [0,1]$,*

$$\phi(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq t\phi(\boldsymbol{x}) + (1-t)\phi(\boldsymbol{y}) - \frac{\mu}{2}t(1-t)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

2. *For differentiable ϕ, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$,*

$$\phi(\boldsymbol{x}) \geq \phi(\boldsymbol{y}) + \langle \phi'(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2. \tag{4.3}$$

3. *For twice differentiable ϕ, for all $\boldsymbol{x} \in \mathcal{A}$,*

$$\nabla^2 f(\boldsymbol{x}) \succeq \mu \boldsymbol{I}.$$

The first definition is reminiscent of Definition 4.1.2. The second definition strengthens Eq. (4.2) to a *quadratic* lower bound, as illustrated in Fig. 4.2. In both cases setting $\mu = 0$ recovers the corresponding convex formulations. The third definition can be interpreted as saying that the change in the gradient is increasing at least at rate μ.

Example 4.1.7 (Strongly convex functions)**.** *The quadratic function $\boldsymbol{x}^2$ from Example 4.1.4 is 1-strongly convex. If the minimum eigenvalue of $\boldsymbol{A}$, which we denote $\lambda_{\min}$, is positive, then the quadratic form $\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ is $\lambda_{\min}$-strongly convex. In particular, if $\boldsymbol{A} = \lambda \boldsymbol{I}$, then $\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2$ is λ-strongly convex. None of the other functions listed in Example 4.1.4 are strongly convex. However, if we restrict their domain, many of these functions will be strongly convex. A notable example is the exponential function, which, if restricted to $[a, \infty)$ for any $a \in \mathbb{R}$, is $\exp(a)$-strongly convex.*

Exercise 4.1.3 (Checking that functions are strongly convex)**.** *Verify using any of the equivalent definitions that the quadratic function and quadratic form are strongly convex, and the exponential function restricted to $[a, \infty)$ is strongly convex.*

4.2 Properties of Convex Functions

Convex and strongly convex functions have a large number of useful properties. We describe just a few relevant ones here. The first two concern the minima of convex functions.

Proposition 4.2.1. *Assume that a convex function ϕ has a* **local optimum** $\boldsymbol{x}_\star$*, in the sense that for some $\delta > 0$ and all $\boldsymbol{x} \neq \boldsymbol{x}_\star$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\star\|_2 < \delta$, we have $\phi(\boldsymbol{x}) > \phi(\boldsymbol{x}_\star)$. Then if $\boldsymbol{x}'_\star$ is another local optima, $\phi(\boldsymbol{x}_\star) = \phi(\boldsymbol{x}'_\star)$. Hence, $\boldsymbol{x}_\star$ is also a* **global optimum**.

Proof. Assume for purposes of contradiction that $\phi(\boldsymbol{x}_\star) > \phi(\boldsymbol{x}'_\star)$. For $t \in (0, 1)$ sufficiently small, $\boldsymbol{x}_t := (1-t)\boldsymbol{x}_\star + t\boldsymbol{x}'_\star$ satisfies $\|\boldsymbol{x}_t - \boldsymbol{x}_\star\|_2 < \delta$. Then by the definition of convexity and the assumption that $\phi(\boldsymbol{x}_\star) > \phi(\boldsymbol{x}'_\star)$, we have

$$\phi(\boldsymbol{x}_t) \leq (1-t)\phi(\boldsymbol{x}_\star) + t\phi(\boldsymbol{x}'_\star) < \phi(\boldsymbol{x}_\star).$$

But this contradicts the hypothesis that $\phi(\boldsymbol{x}) > \phi(\boldsymbol{x}_\star)$ whenever $\|\boldsymbol{x} - \boldsymbol{x}_\star\|_2 < \delta$, a contradiction. Hence $\phi(\boldsymbol{x}_\star) \leq \phi(\boldsymbol{x}'_\star)$. Applying the same reasoning with roles of $\boldsymbol{x}_\star$ and $\boldsymbol{x}'_\star$ reversed, we conclude that $\phi(\boldsymbol{x}_\star) \geq \phi(\boldsymbol{x}'_\star)$, so in fact we must have $\phi(\boldsymbol{x}_\star) = \phi(\boldsymbol{x}'_\star)$. □

Proposition 4.2.2. *Assume that a strictly convex function ϕ has a* **local optimum** $\boldsymbol{x}_\star$*. Then for $\boldsymbol{x} \in \mathcal{A}$ with $\boldsymbol{x} \neq \boldsymbol{x}_\star$, we have $\phi(\boldsymbol{x}) > \phi(\boldsymbol{x}_\star)$. Hence, $\boldsymbol{x}_\star$ is the* unique *global optimum.*

Exercise 4.2.1. *Prove Proposition 4.2.2.*

The first one provides a probabilistic interpretation of convexity, which we can motivate as follows. For $t \in [0, 1]$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$, define $\boldsymbol{X}_{t,\boldsymbol{x},\boldsymbol{y}}$ to be a random vector which takes the value $\boldsymbol{x}$ with probability t and the value

$\boldsymbol{y}$ with probability $1 - t$, the definition of convexity can be rewritten as requiring that, for all $t, \boldsymbol{x}, \boldsymbol{y}$, $\phi(\mathbb{E}\{\boldsymbol{X}_{t,\boldsymbol{x},\boldsymbol{y}}\}) \leq \mathbb{E}\{\phi(\boldsymbol{X}_{t,\boldsymbol{x},\boldsymbol{y}})\}$. This inequality actually holds for all random variables when ϕ is convex:

Lemma 4.2.3 (Jensen's inequality)**.** *The function $\phi\colon \mathcal{A} \to \mathbb{R}$ is convex if and only if, for every random vector defined on $\mathcal{A}$,*

$$\phi(\mathbb{E}\{\boldsymbol{X}\}) \leq \mathbb{E}\{\phi(\boldsymbol{X})\}.$$

If ϕ is concave, the direction of the inequality is reversed, so $\phi(\mathbb{E}\{\boldsymbol{X}\}) \geq \mathbb{E}\{\phi(\boldsymbol{X})\}$.

Strongly convex functions also satisfy another quadratic lower bound.

Lemma 4.2.4. *If ϕ is μ-strongly convex, then*

$$\langle \phi'(\boldsymbol{x}) - \phi'(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \mu \left\| \boldsymbol{x} - \boldsymbol{y} \right\|_2^2. \tag{4.4}$$

Proof. Add Eq. (4.3) to a copy of itself with $\boldsymbol{x}$ and $\boldsymbol{y}$ reversed in the copy. □

Finally, we would like to be able to construct new convex functions using other convex (and sometimes non-convex) functions. First, we have some operations involving convex and linear functions.

Lemma 4.2.5 (Operations on convex functions)**.**

1. *If ϕ_i is μ_i-strongly convex, then for $a_i > 0$, the function $\phi = a_1\phi_1 + a_2\phi_2$ is $(a_1\mu_1 + a_2\mu_2)$-strongly convex.*
2. *If $\phi : \mathbb{R}^{D'} \to \mathbb{R}$ is convex, then for any $\boldsymbol{A} \in \mathbb{R}^{D' \times D}$ and $\boldsymbol{b} \in \mathbb{R}^{D'}$, the function $\boldsymbol{x} \mapsto \phi(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ is convex.*

The composition of two convex functions is not necessarily convex, as one must be careful with the monotonicity properties of the outer function.

Lemma 4.2.6 (Composition of functions)**.** *If $\phi = \phi_1 \circ \phi_2$, then ϕ is convex if*

1. *ϕ_2 is convex and ϕ_1 is convex and non-decreasing, or*
2. *ϕ_2 is concave and ϕ_1 is convex and non-increasing.*

Exercise 4.2.2 (Operations on convex functions)**.** *Prove Lemma 4.2.5 under the assumption that all functions are differentiable.*

Exercise 4.2.3 (Composition of functions)**.** *Prove Lemma 4.2.6 under the assumption that all functions are differentiable.*

4.3 Other Regularity Conditions

An important property that is "dual" to strong convexity is that of strong smoothness.

Definition 4.3.1 (Strong smoothness)**.** *A function ϕ is* L**-strongly smooth** *if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$,*

$$\|\phi'(\boldsymbol{x}) - \phi'(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

A function ϕ being L-strongly smooth can be interpreted as meaning that the gradient of ϕ does not change faster than rate L. Hence, if it exists, the operator norm of the Hessian matrix is bounded.

Proposition 4.3.2. *If ϕ is twice-differentiable and $L := \sup_{\boldsymbol{x}\in\mathcal{A}}\|\phi''(\boldsymbol{x})\|_2 < \infty$, then ϕ is L-strongly smooth.*

Exercise 4.3.1. *Prove Proposition 4.3.2 using the fundamental theorem of calculus for line integrals.*

If a function is also convex, then it has a property known as co-coercivity:

Lemma 4.3.3. *If $\phi\colon \mathcal{A} \to \mathbb{R}$ is convex and L-strongly smooth, then ϕ' is* L**-co-coercive***: for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$,*

$$\|\phi'(\boldsymbol{x}) - \phi'(\boldsymbol{y})\|_2^2 \leq L\langle\phi'(\boldsymbol{x}) - \phi'(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle.$$

If a function ϕ is strongly smooth and convex, then bounds like those for strong convexity (specifically, Eqs. (4.3) and (4.4)) – but in the opposite direction – hold.

Lemma 4.3.4. *If $\phi\colon \mathcal{A} \to \mathbb{R}$ is convex and L-strongly smooth, then for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$,*

$$0 \leq \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \phi'(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

and hence

$$\langle \phi'(\boldsymbol{x}) - \phi'(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

Thus, we conclude that if a function is L-strongly smooth and μ-strongly convex, $L \geq \mu$. The ratio $\rho := L/\mu$ is called the ***condition number***.

Example 4.3.5 (Regression and classification)**.** *The regression and classification models described in Examples 1.1.1 to 1.1.3 are all convex, and are λ-strongly convex due to the $\lambda\|\boldsymbol{x}\|_2^2$ regulation term. For logistic regression and SVMs, the data-dependent term of the loss (or, equivalently, the loss with $\lambda = 0$) is not strongly convex. For linear regression, the data-dependent term of the loss is strongly convex if the sample covariance $\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{z}_n \boldsymbol{z}_n^\top$ is positive-definite, in which case the strong convexity constant is equal its smallest eigenvalue.*

4.4 Error of Taylor Series Approximations

We have now seen how convex functions are closely related to bounds on first-order Taylor approximations while strongly convex functions are closely related to bounds on second-order Taylor approximations. Thus, Taylor approximation arguments are often used in combination with convex analysis. For this approach to be fruitful, we must make use of the fact that error of the approximation has an explicit form that, under smoothness assumptions, can be bounded. We will consider only univariate functions. Extensions to multivariate functions, while relatively straightforward, result in much less transparent expressions.

Theorem 4.4.1 (Taylor's theorem, Lagrange and integral remainder)**.** *If $\phi\colon [a, b] \to \mathbb{R}$ is a p-times continuously differentiable function, then for any $x, u \in [a, b]$,*

$$\phi(x) = \sum_{i=0}^{p-1} \frac{\phi^{(i)}(u)}{i!}(x - u)^i + R_p,$$

where $\phi^{(p)}$ denotes the pth derivative of ϕ and, for some $v \in [x, u]$,

$$R_p = \frac{\phi^{(p)}(v)}{p!}(x-u)^p.$$

Alternatively,

$$R_p = \int_u^x \frac{\phi^{(p)}(t)}{(p-1)!}(x-t)^{p-1}\mathrm{d}t.$$

As an example application of Theorem 4.4.1, the next result provides a bound on the expected error for a random function.

Corollary 4.4.2. *Under the conditions of Theorem 4.4.1, if ϕ is a random function satisfying $\mathbb{E}\{|\phi^{(p)}(v)|\} \leq M$ for all $v \in [x, u]$, then*

$$\mathbb{E}\{|R_p|\} \leq \frac{M}{p!}|x-u|^p.$$

Proof. Using the integral remainder formula, Exercise 2.4.5 and Theorem 2.4.6, and the assumption that $\mathbb{E}\{|\phi^{(p)}(v)|\} \leq M$, we have

$$\begin{aligned}
\mathbb{E}\{|R_p|\} &= \mathbb{E}\left\{\left|\int_u^x \frac{\phi^{(p)}(t)}{(p-1)!}(x-t)^{p-1}\mathrm{d}t\right|\right\} \\
&\leq \mathbb{E}\left\{\int_x^u \frac{|\phi^{(p)}(t)|}{(p-1)!}|x-t|^{p-1}\mathrm{d}t\right\} \\
&= \int_x^u \frac{\mathbb{E}\{|\phi^{(p)}(t)|\}}{(p-1)!}|x-t|^{p-1}\mathrm{d}t \\
&\leq \frac{M}{(p-1)!}\int_x^u |x-t|^{p-1}\mathrm{d}t \\
&= \frac{M}{p!}|x-u|^p.
\end{aligned}$$

□

4.5 Error Analysis of Stochastic Gradient Descent

We now turn to applying the results from this chapter to the error analysis of stochastic gradient descent (SGD), which is described in Section 1.1. Our

results will apply beyond the finite-sum optimization setting described there, so we will adjust our notation slightly while keeping in mind that special case. We will denote the function we wish to minimize as f, so our goal is to estimate the minimizer

$$\boldsymbol{x}_\star = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}).$$

We assume access to a sequence of independent, unbiased stochastic gradient estimates $\widehat{f}'_1, \widehat{f}'_2, \ldots$ such that for any $\boldsymbol{x} \in \mathcal{A}$, we have $f'(\boldsymbol{x}) = \mathbb{E}\{\widehat{f}'_k(\boldsymbol{x})\}$. So, in the notation of Section 1.1, we could have $f = \mathcal{L}$ and $\widehat{f}'_k = \mathcal{L}'_k$. The SGD algorithm update is therefore given by

$$\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k - \eta_{k+1} \widehat{f}'_{k+1}(\boldsymbol{x}_k), \tag{4.5}$$

where $\{\eta_k\}_{k \in \mathbb{N}}$ is a sequence of positive step sizes. We use $\mathbb{E}_k$ as a shorthand for the conditional expectation $\mathbb{E}(\cdot \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_k, \widehat{f}'_1, \ldots, \widehat{f}'_k)$. The following assumption formalizes the independence and unbiasedness requirements for the gradient estimates:

Assumption 4.5.1. *The sequence $(\widehat{f}'_k)_{k \in \mathbb{N}}$ is i.i.d. and for all $k \in \mathbb{N}$ and $\boldsymbol{x} \in \mathcal{A}$, $f'(\boldsymbol{x}) = \mathbb{E}_{k-1}\{\widehat{f}'_k(\boldsymbol{x})\}$.*

Define the ***stochastic gradient error*** by $\varepsilon_k := \widehat{f}'_k - f'$, which by Assumption 4.5.1 satisfies $\mathbb{E}_{k-1}(\varepsilon_k) = 0$. Hence, the one-step update in Eq. (4.5) can be rewritten as

$$\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k - \eta_{k+1} \{f'(\boldsymbol{x}_k) + \varepsilon_{k+1}(\boldsymbol{x}_k)\}.$$

To keep the analyses in this chapter as simple as possible, we focus on the case of strongly convex functions:

Assumption 4.5.2. *There exists $\mu > 0$ such that the function f is μ-strongly convex.*

We also require an assumption about the behavior of the stochastic gradients, namely that they are co-coercive.

Assumption 4.5.3. *There exists $L > 0$ such that for all $k \in \mathbb{N}$, $\widehat{f}'_k$ is L-co-coercive.*

We will also assume that the noise variance at the optimum is finite:

Assumption 4.5.4. *There exists $\sigma > 0$ such that for all $k \in \mathbb{N}$,*

$$\mathbb{E}_{k-1}\left\{\|\widehat{f}_k'(\boldsymbol{x}_\star)\|_2^2\right\} \leq \sigma^2.$$

Often optimization error guarantees require the gradient error to be uniformly bounded. However, this is a very strong assumption that rarely holds in practice. Assuming bounded variance at the optimum is a much weaker condition.

The analyses in this section will concern the squared error $E_k := \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2$. We start with a useful lemma, which we will use for our error analysis assuming either a fixed or decreasing step size.

Lemma 4.5.5. *If Assumptions 4.5.1 to 4.5.4 hold and $\eta_k \in (0, 1/L)$, then*

$$\mathbb{E}_{k-1}(E_k) \leq \{1 - 2\eta_k\mu(1 - \eta_k L)\}E_{k-1} + 2\eta_k^2\sigma^2. \tag{4.6}$$

Proof of Lemma 4.5.5. We have

$$\begin{aligned}
\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2 &\overset{(i)}{=} \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2 - 2\eta_k\langle \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star, \widehat{f}_k'(\boldsymbol{x}_{k-1})\rangle + \eta_k^2\|\widehat{f}_k'(\boldsymbol{x}_{k-1})\|_2^2 \\
&= \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2 - 2\eta_k\langle \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star, \widehat{f}_k'(\boldsymbol{x}_{k-1})\rangle \\
&\quad + \eta_k^2\|\widehat{f}_k'(\boldsymbol{x}_{k-1}) - \widehat{f}_k'(\boldsymbol{x}_\star) + \widehat{f}_k'(\boldsymbol{x}_\star)\|_2^2 \\
&\overset{(ii)}{\leq} \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2 - 2\eta_k\langle \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star, \widehat{f}_k'(\boldsymbol{x}_{k-1})\rangle \\
&\quad + 2\eta_k^2\|\widehat{f}_k'(\boldsymbol{x}_{k-1}) - \widehat{f}_k'(\boldsymbol{x}_\star)\|_2^2 + 2\eta_k^2\|\widehat{f}_k'(\boldsymbol{x}_\star)\|_2^2 \\
&\overset{(iii)}{\leq} \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2 - 2\eta_k\langle \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star, \widehat{f}_k'(\boldsymbol{x}_{k-1})\rangle \\
&\quad + 2\eta_k^2 L\langle \widehat{f}_k'(\boldsymbol{x}_{k-1}) - \widehat{f}_k'(\boldsymbol{x}_\star), \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\rangle + 2\eta_k^2\|\widehat{f}_k'(\boldsymbol{x}_\star)\|_2^2, \qquad (4.7)
\end{aligned}$$

where (i) follows from Eq. (4.5) and expanding the norm, (ii) follows from the triangle inequality, and (iii) follows from using Assumption 4.5.3 to bound the penultimate term. Next, take the conditional expectation of Eq. (4.7), apply Assumptions 4.5.1 and 4.5.4, and then use Eq. (4.4) (since Assumption 4.5.2 holds):

$$\begin{aligned}
&\mathbb{E}_{k-1}(E_k) \\
&\leq \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2 - 2\eta_k\langle \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star, \mathbb{E}_{k-1}\{\widehat{f}_k'(\boldsymbol{x}_{k-1})\}\rangle \\
&\quad + 2\eta_k^2 L\langle \mathbb{E}_{k-1}\{\widehat{f}_k'(\boldsymbol{x}_{k-1}) - \widehat{f}_k'(\boldsymbol{x}_\star)\}, \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\rangle + 2\eta_k^2\mathbb{E}_{k-1}\{\|\widehat{f}_k'(\boldsymbol{x}_\star)\|_2^2\} \\
&\leq \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2 - 2\eta_k\langle \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star, f'(\boldsymbol{x}_{k-1})\rangle
\end{aligned}$$

$$
\begin{aligned}
&\quad + 2\eta_k^2 L\langle f'(\boldsymbol{x}_{k-1}) - f'(\boldsymbol{x}_\star), \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\rangle + 2\eta_k^2\sigma^2 \\
&= \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2 - 2\eta_k(1-\eta_k L)\langle \boldsymbol{x}_{k-1} - \boldsymbol{x}_\star, f'(\boldsymbol{x}_{k-1}) - f'(\boldsymbol{x}_\star)\rangle + 2\eta_k^2\sigma^2 \\
&\leq \{1 - 2\eta_k\mu(1-\eta_k L)\}\|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2 + 2\eta_k^2\sigma^2.
\end{aligned}
$$

□

4.6 Error Analysis of SGD with Constant Step Size

We are now ready to analyze the behavior of SGD with *constant* step size $\eta_k = \eta$. For SGD with constant step size, an inductive argument yields the following error bound on the individual iterates:

Theorem 4.6.1. *If Assumptions 4.5.1 to 4.5.4 hold and $\eta_k = \eta \in (0, \frac{1}{2L})$, then for $\beta := 1 - 2\eta\mu(1-\eta L)$,*

$$
\mathbb{E}(E_k) \leq \beta^k \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 + \frac{2\eta}{\mu}\sigma^2, \tag{4.8}
$$

and therefore

$$
\mathbb{E}(\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2) \leq \beta^{k/2}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2 + \frac{2^{1/2}\eta^{1/2}}{\mu^{1/2}}\sigma.
$$

Exercise 4.6.1. *Show that for nonnegative constants $a, b > 0$,*

$$
\sqrt{a^2 + b^2} \leq a + b.
$$

Exercise 4.6.2 (Error of SGD with constant step size)**.** *Prove Theorem 4.6.1. [Hint: Using the geometric series formula $\sum_{\ell=0}^{k-1}\alpha^\ell = (1-\alpha^k)/(1-\alpha)$ will be helpful for proving the first inequality. The bound from Exercise 4.6.1 will helpful for proving the second inequality.]*

Theorem 4.6.1 provides a few useful insights. First, the only dependence on the initial condition is through the first term, which decays exponentially fast. Hence, SGD with constant step size has only very weak dependence on $\boldsymbol{x}_0$. On the other hand, the second term indicates an irreducible expected

squared error of order η. This irreducible error is scaled by the noise level σ^2: as we might expect, noisier stochastic gradients leads to less accurate estimates of $\boldsymbol{x}_\star$. The irreducible error is also scale by the inverse of the strong convexity constant μ. A smaller μ implies a the flatter function, which leads to larger error in $\boldsymbol{x}_\star$.

In Chapter 1 we saw that iterate averaging can provide much smaller error than individual iterates. Generalizing the earlier definition of the iterate average, for $k_0 \leq k$, the ***iterate average from*** k_0 ***to*** $k-1$ is given by

$$\bar{\boldsymbol{x}}_{k_0:k} := \frac{1}{k-k_0} \sum_{\ell=k_0}^{k-1} \boldsymbol{x}_\ell.$$

We now show that, in fact, $\bar{\boldsymbol{x}}_{0:k}$ has error of order η, which is superior to a single iterate as long as η is fairly small (which is typically the case). The result requires an additional smoothness assumption on the stochastic gradients.

Assumption 4.6.2. *There for the same $L > 0$ as Assumption 4.5.3 and a constant $M > 0$, for all $k \in \mathbb{N}$ and $\boldsymbol{x} \in \mathcal{A}$,*

$$\mathbb{E}_{k-1}\{\|\widehat{f}_k''(\boldsymbol{x})\|_2\} \leq L \qquad \text{and} \qquad \mathbb{E}_{k-1}\{\|\widehat{f}_k'''(\boldsymbol{x})\|_2\} \leq M.$$

Recall that $\rho := L/\mu$ is the condition number of f.

Theorem 4.6.3. *If Assumptions 4.5.1 to 4.6.2 hold and $\eta_k = \eta \in (0, 1/L)$, then for $\varrho := M/\mu$,*

$$\begin{aligned}
&\mathbb{E}(\|\bar{\boldsymbol{x}}_{0:k} - \boldsymbol{x}_\star\|_2) \\
&\quad \leq \frac{\varrho\eta\sigma^2}{\mu} + \frac{\sigma}{\mu k^{1/2}}(1 + 2^{3/2}\rho^{1/2}) \\
&\qquad + \frac{1}{\mu\eta k}\left\{\frac{\varrho}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 + 2(1+\rho^{1/2})\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2 + 2^{1/2}\rho^{1/2}\sigma\right\}.
\end{aligned}$$

Let us interpret the bound, treating the step size as fixed. The first term in the bound is the irreducible error, which is of order η, an improvement over the single-iterate bound. The second term is the error due to the Monte Carlo (stochastic gradient) noise, which has the standard Monte Carlo rate of $1/k^{1/2}$. The final term depends on the initialization, but decays at the faster rate of $1/(\eta k)$. Since this error is primarily deterministic, it depends on the "distance traveled" after k steps, which is of order ηk. To improve the dependence on the initialization by making is decrease exponentially fast, we can start to average after $k/2$ iterations:

Corollary 4.6.4. *If Assumptions 4.5.1 to 4.6.2 hold and $\eta_k = \eta \in (0, \frac{1}{2L})$, then for k even, $\varrho := M/\mu$, and $\beta := 1 - 2\eta\mu(1 - \eta L)$,*

$$\begin{aligned}(4.9)\qquad &\mathbb{E}\big\{\|\bar{\boldsymbol{x}}_{k/2:k} - \boldsymbol{x}_\star\|_2\big\} \\ &\le \frac{\varrho\eta\sigma^2}{\mu} + \frac{\sigma}{\mu k^{1/2}}\Big(1 + 2^{3/2}\rho^{1/2}\Big) + \frac{2\varrho\sigma^2}{\mu^2 k} + \frac{3\rho^{1/2}\sigma}{\mu\eta k}(2 + \rho^{1/2}) \\ &\quad + \beta^{k/2}\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2}{\mu\eta k}\bigg\{2(1 + \rho^{1/2}) + \frac{\varrho}{2\mu}\beta^{k/2}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2\bigg\}.\end{aligned}$$

Exercise 4.6.3 (Error of iterate averaging with constant step size)**.** *Use Theorems 4.6.1 and 4.6.3 to prove Corollary 4.6.4.*

Proof of Theorem 4.6.3. The proof is based essentially on a Taylor series approximation argument, which starts with the observation that, by Theorem 4.4.1, $\widehat{f}'_k(\boldsymbol{x}_{k-1}) \approx \widehat{f}'_k(\boldsymbol{x}_\star) + f''(\boldsymbol{x}_\star)(\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star) + r_k$, where r_k is a remainder term. Rearranging and averaging, we obtain

$$\frac{1}{k}\sum_{\ell=1}^{k} f''(\boldsymbol{x}_\star)(\boldsymbol{x}_{\ell-1} - \boldsymbol{x}_\star) \approx \frac{1}{k}\sum_{\ell=1}^{k}\{\widehat{f}'_\ell(\boldsymbol{x}_{\ell-1}) - \widehat{f}'_\ell(\boldsymbol{x}_\star) - r_\ell\}.$$

Since $\widehat{f}'_1(\boldsymbol{x}_\star), \ldots, \widehat{f}'_k(\boldsymbol{x}_\star)$ is essentially an i.i.d. sequence with mean zero, its average is of order $1/k$. Using Eq. (4.5), we have $\widehat{f}'_k(\boldsymbol{x}_{k-1}) = (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k)/\eta$, so

$$(4.10)\qquad \frac{1}{k}\sum_{\ell=1}^{k}\widehat{f}'_\ell(\boldsymbol{x}_{\ell-1}) = \frac{1}{\eta k}(\boldsymbol{x}_0 - \boldsymbol{x}_{k+1}),$$

which is also of order $1/k$. Finally the average of the remainder terms r_k is of order η.

To make these arguments precise, we have

$$\begin{aligned}f''(\boldsymbol{x}_\star)(\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star) &= \widehat{f}''_k(\boldsymbol{x}_\star)(\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star) + \{f''(\boldsymbol{x}_\star) - \widehat{f}''_k(\boldsymbol{x}_\star)\}(\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star) \\ &= \underbrace{\widehat{f}'_k(\boldsymbol{x}_{k-1})}_{\mathrm{I}_k} - \underbrace{\widehat{f}'_k(\boldsymbol{x}_\star)}_{\mathrm{II}_k} + \underbrace{\widehat{f}''_k(\boldsymbol{x}_\star)(\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star) - \widehat{f}'_k(\boldsymbol{x}_{k-1}) + \widehat{f}'_k(\boldsymbol{x}_\star)}_{\mathrm{III}_k} \\ &\quad + \underbrace{\{f''(\boldsymbol{x}_\star) - \widehat{f}''_k(\boldsymbol{x}_\star)\}(\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star)}_{\mathrm{IV}_k}.\end{aligned}$$

After summing both sides, we first apply the triangle inequality, and then use the definition of the spectral norm (see Appendix A.2):

$$\begin{aligned}\left\|\frac{1}{k}\sum_{\ell=0}^{k-1}\boldsymbol{x}_\ell-\boldsymbol{x}_\star\right\|_2 &\le \left\|f''(\boldsymbol{x}_\star)^{-1}\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{I}_\ell\right\|_2+\left\|f''(\boldsymbol{x}_\star)^{-1}\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{II}_\ell\right\|_2\\ &\quad+\left\|f''(\boldsymbol{x}_\star)^{-1}\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{III}_\ell\right\|_2+\left\|f''(\boldsymbol{x}_\star)^{-1}\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{IV}_\ell\right\|_2\\ &\le \|f''(\boldsymbol{x}_\star)^{-1}\|_2\left\|\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{I}_\ell\right\|_2+\|f''(\boldsymbol{x}_\star)^{-1}\|_2\left\|\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{II}_\ell\right\|_2\\ &\quad+\|f''(\boldsymbol{x}_\star)^{-1}\|_2\left\|\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{III}_\ell\right\|_2+\|f''(\boldsymbol{x}_\star)^{-1}\|_2\left\|\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{IV}_\ell\right\|_2.\end{aligned}$$

Next, we take the expectations of both sides and bound each term in turn. First, using Eq. (4.10), the triangle inequality, and Theorem 4.6.1, we have

$$\begin{aligned}&\mathbb{E}\Big(\|\tfrac{1}{k}\textstyle\sum_{\ell=0}^{k-1}\mathrm{I}_\ell\|_2\Big)\\ &\le\frac{1}{\eta k}\{\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2+\mathbb{E}(\|\boldsymbol{x}_k-\boldsymbol{x}_\star\|_2)\}\\ &\le\frac{1}{\eta k}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2+\frac{1}{\eta k}\{1-2\eta\mu(1-\eta L)\}^{k/2}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2+\frac{1}{\eta k}\frac{2^{1/2}\eta^{1/2}\sigma}{\mu^{1/2}}\\ &\le\frac{2}{\eta k}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2+\frac{2^{1/2}\sigma}{\eta^{1/2}\mu^{1/2}k}.\end{aligned}$$

Using Assumptions 4.5.1 and 4.5.4 and Jensen's inequality, the second term can be bounded as

$$\mathbb{E}\Big(\|\tfrac{1}{k}\textstyle\sum_{\ell=0}^{k-1}\mathrm{II}_\ell\|_2\Big)\le\frac{\sigma}{k^{1/2}}.$$

By Corollary 4.4.2 and Assumption 4.6.2,

$$\mathbb{E}_{k-1}(|\mathrm{III}_k|)\le\frac{M\|\boldsymbol{x}_{k-1}-\boldsymbol{x}_\star\|_2^2}{2}.$$

Combining this inequality with Assumption 4.6.2 and Theorem 4.6.1, we

have

$$\begin{aligned}\mathbb{E}\left(\left\|\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{III}_\ell\right\|_2\right) &\le \frac{M}{2}\frac{1}{k}\sum_{\ell=0}^{k-1}\mathbb{E}(\|\boldsymbol{x}_\ell-\boldsymbol{x}_\star\|_2^2)\\ &\le \frac{M}{2}\frac{1}{k}\sum_{\ell=0}^{k-1}\{1-2\eta(1-\eta L)\}^\ell\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2^2+\frac{M\eta\sigma^2}{\mu}\\ &\le \frac{M}{2}\frac{1}{k}\sum_{\ell=0}^{k-1}\{1-2\eta(1-\eta L)\}^\ell\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2^2+\frac{M\eta\sigma^2}{\mu}\\ &\le \frac{M}{2\eta\mu k}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2^2+\frac{M\eta\sigma^2}{\mu}.\end{aligned}$$

Similarly, for the fourth term,

$$\begin{aligned}\mathbb{E}\left(\left\|\frac{1}{k}\sum_{\ell=0}^{k-1}\mathrm{IV}_\ell\right\|_2\right) &\le \frac{2L}{k}\left\{\sum_{\ell=0}^{k-1}\mathbb{E}(\|\boldsymbol{x}_\ell-\boldsymbol{x}_\star\|_2^2)\right\}^{1/2}\\ &\le \frac{2L}{k^{1/2}}\left\{\frac{1}{\eta\mu k}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2^2+\frac{2\eta}{\mu}\sigma^2\right\}^{1/2}\\ &\le \frac{2L}{\eta^{1/2}\mu^{1/2}k}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2+\frac{2^{3/2}L\eta^{1/2}\sigma}{\mu^{1/2}k^{1/2}}.\end{aligned} \tag{4.11}$$

Since f is strongly convex, all the eigenvalues of $f''(\boldsymbol{x}_\star)$ are positive. So $\|f''(\boldsymbol{x}_\star)^{-1}\|_2$ is equal to the reciprocal of the smallest eigenvalue, which is upper bounded by $1/\mu$. So, putting everything together, we have

$$\begin{aligned}\mu\mathbb{E}(\|\bar{\boldsymbol{x}}_{0:k}-\boldsymbol{x}_\star\|_2) &\le \frac{2}{\eta\mu k}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2+\frac{2^{1/2}\sigma}{\eta^{1/2}\mu^{1/2}k}+\frac{\sigma}{k^{1/2}}\\ &\quad+\frac{M}{2\eta\mu k}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2^2+\frac{M\eta\sigma^2}{\mu}\\ &\quad+\frac{2L}{\eta^{1/2}\mu^{1/2}k}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2+\frac{2^{3/2}L\eta^{1/2}\sigma}{\mu^{1/2}k^{1/2}}\\ \mathbb{E}(\|\bar{\boldsymbol{x}}_{0:k}-\boldsymbol{x}_\star\|_2) &\le \frac{\varrho\eta\sigma^2}{\mu}+\frac{\sigma}{\mu k^{1/2}}(1+2^{3/2}\rho^{1/2})\\ &\quad+\frac{1}{\mu\eta k}\Big\{\frac{\varrho}{2}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2^2+2(1+\rho^{1/2})\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2+2^{1/2}\rho^{1/2}\sigma\Big\},\end{aligned}$$

where to obtain the last inequality we have used the assumption that $\eta \le L$ and the definition $\rho = L/\mu$. □

Exercise 4.6.4 (Bounding the Hessian error)**.** *Prove Eq.* (4.11)*.* [Hint: use the fact that $f''(\boldsymbol{x}_\star) = \mathbb{E}\{\tilde{f}_k''(\boldsymbol{x}_\star)\}$, where $\tilde{f}_k''$ is an independent copy of $\widehat{f}_k''$.]

4.7 Convergence of SGD with Decreasing Step Size

We now consider the case when $\eta_k \to 0$ for $k \to \infty$ and wish to show that, in fact, $\boldsymbol{x}_k$ converges to the optimum $\boldsymbol{x}_\star = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$. Specifically, we will focus on the widely used step size schedule $\eta_k = \eta/k^\alpha$. The main result is the following quantitative bound on $\mathbb{E}(E_k)$:

Theorem 4.7.1. *Under Assumptions 4.5.1 to 4.5.4, if $\eta_k = \eta/k^\alpha$ for $\alpha \in (1/2, 1)$ and $\eta \in (0, \frac{1}{2L})$, then the iterates of Eq.* (4.5) *satisfy*

$$\mathbb{E}(E_k) \le \exp\Big\{-\frac{\mu\eta}{2}(k^{1-\alpha} - 1)\Big\}\left(\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 + \frac{4\alpha\sigma^2\eta^2}{2\alpha - 1}\right) + \frac{2\sigma^2\eta}{\mu k^\alpha}.$$

The term $\exp(-\frac{\mu\eta}{2}k^{1-\alpha})\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2$ is "transient": it decays sub-exponentially fast but depends on the (expected) initial squared distance to the optimum, $\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2$. The (asymptotically relevant) slowly decaying term $\frac{2\sigma^2\eta}{\mu k^\alpha}$ does not depend on the initial conditions. Thus, taking $\alpha \to 1$ results in a faster convergence rate; however, it also results in the transient term no longer converging to zero as $k \to \infty$. Exercise 4.7.4 asks you to derive a bound on $\mathbb{E}(E_k)$ that does converge to zero.

Proof of Theorem 4.7.1. Using Lemma 4.5.5 and the bound $1 - \eta_k L \ge 1/2$, we have the recursion

$$\mathbb{E}(E_k) \le (1 - \mu\eta_k)\|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2 + 2\sigma^2\eta_k^2. \tag{4.12}$$

Applying Lemma 4.7.2 with $c = 2\sigma^2$ and $m = k/2$ and using the bounds

$$\sum_{\ell=k/2+1}^{k} \eta_\ell \ge \eta\frac{k^{1-\alpha} - 1}{2} \quad \text{and} \quad \sum_{\ell=1}^{k/2} \eta_\ell^2 \le \eta^2\frac{2\alpha}{2\alpha - 1} \tag{4.13}$$

yields the bound

$$\begin{aligned}\mathbb{E}(E_k) &\le \exp\left\{-\mu\sum_{\ell=1}^{k}\eta_\ell\right\}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2^2 + 2\sigma^2\exp\left\{-\mu\sum_{\ell=k/2+1}^{k}\eta_\ell\right\}\left(\sum_{\ell=1}^{k/2}\eta_\ell^2\right) + \frac{2\sigma^2\eta_k}{\mu}\\ &\le \exp\left\{-\mu\sum_{\ell=k/2+1}^{k}\eta_\ell\right\}\left(\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2^2 + 2\sigma^2\sum_{\ell=1}^{k/2}\eta_\ell^2\right) + \frac{2\sigma^2\eta_k}{\mu}\\ &\le \exp\Big\{-\frac{\mu\eta}{2}(k^{1-\alpha}-1)\Big\}\left(\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|_2^2 + \frac{4\alpha\sigma^2\eta^2}{2\alpha-1}\right) + \frac{2\sigma^2\eta}{\mu k^\alpha}.\end{aligned}$$

□

Lemma 4.7.2. *For $\mu > 0$ and non-increasing $\eta_k > 0$ with $\eta_0 < 1/\mu$, if the sequence $\delta_1, \delta_2, \ldots$ satisfies*

$$\delta_k \le (1-\mu\eta_k)\delta_{k-1} + c\eta_k^2,$$

then for any $m \in [k]$,

$$\delta_k \le \exp\left\{-\mu\sum_{\ell=1}^{k}\eta_\ell\right\}\delta_0 + c\exp\left\{-\mu\sum_{\ell=m+1}^{k}\eta_\ell\right\}\left(\sum_{\ell=1}^{m}\eta_\ell^2\right) + \frac{c\,\eta_m}{\mu}. \tag{4.14}$$

Proof. By induction,

$$\delta_k \le \prod_{\ell=1}^{k}(1-\mu\eta_\ell)\delta_0 + c\sum_{i=1}^{k}\eta_i^2\prod_{\ell=i+1}^{k}(1-\mu\eta_\ell). \tag{4.15}$$

For the first term, using the basic inequality that $1-a \le \exp(-a)$ (for $a \in \mathbb{R}$), we have

$$\prod_{\ell=1}^{k}(1-\mu\eta_\ell)\delta_0 \le \exp\left(-\sum_{\ell=1}^{k}\mu\eta_\ell\right)\delta_0.$$

For the second term, clearly for any $m \in [k]$,

$$\sum_{i=1}^{k}\eta_i^2\prod_{\ell=i+1}^{k}(1-\mu\eta_\ell) = \underbrace{\sum_{i=1}^{m}\eta_i^2\prod_{\ell=i+1}^{k}(1-\mu\eta_\ell)}_{\star} + \underbrace{\sum_{i=m+1}^{k}\eta_i^2\prod_{\ell=i+1}^{k}(1-\mu\eta_\ell)}_{\star\star}.$$

Since $\mu > 0$, we then have

$$\star \leq \left(\sum_{i=1}^{m} \eta_i^2\right) \prod_{\ell=m+1}^{k} (1 - \mu\eta_\ell) \leq \left(\sum_{\ell=1}^{m} \eta_\ell^2\right) \exp\left(-\sum_{\ell=m+1}^{k} \mu\eta_\ell\right).$$

Since in addition η_k is non-increasing,

$$\begin{aligned} \star\star &\leq \eta_m \sum_{i=m+1}^{k} \eta_i \prod_{\ell=i+1}^{k} (1 - \mu\eta_\ell) \\ &= \frac{\eta_m}{\mu} \sum_{i=m+1}^{k} \mu\eta_i \prod_{\ell=i+1}^{k} (1 - \mu\eta_\ell) \leq \frac{\eta_m}{\mu}. \end{aligned} \tag{4.16}$$

□

Exercise 4.7.1. *Prove Eq.* (4.15).

Exercise 4.7.2 (Polynomial sum bounds)**.** *Verify the bounds given in Eq.* (4.13). [Hint: bound each sum by an integral]

Exercise 4.7.3 (Stick-breaking inequality)**.** *Verify Eq.* (4.16) *by showing that, for any sequence* $\beta_i \in [0, 1]$*,*

$$\sum_{i=1}^{k} \beta_i \prod_{\ell=i+1}^{k} (1 - \beta_\ell) \leq 1.$$

[Hint: Imagine starting with a stick of length 1 and, at step i ($i = 0, \ldots, k-1$), breaking off and discarding $(100\beta_{k-i})\%$ of the remaining stick.]

Exercise 4.7.4 (SGD error bound when $\alpha = 1$)**.** *Using Lemma 4.7.2 and Eq.* (4.12)*, derive a bound on* $\mathbb{E}(E_k)$ *in the case that* $\alpha = 1$*. The term in the bound depending on* $\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2$ *will be polynomial in* k*; the other term will be of order* $1/k$*.* [Hint: use $m = k$, so the the middle term in the right-hand size of Eq. (4.14) is zero.]

Chapter 5

Metrics for Probability Distributions

With Applications to Convergence of SGD to Stationarity

Introduction to measuring the distance between probability distributions, using Wasserstein distances. As an application, proves convergence to stationarity of SGD with constant step size.

5.1 Metric Convergence

Convergence in distribution is often too weak of a guarantee for practical purposes. If $\boldsymbol{X}_k \xrightarrow{d} \boldsymbol{X}_\infty$ we are only guaranteed that expectations of bounded, continuous functions converge. Moreover, for any particular bounded, continuous ϕ, $\mathbb{E}\{\phi(\boldsymbol{X}_k)\}$ may converge to $\mathbb{E}\{\phi(\boldsymbol{X}_\infty)\}$ arbitrarily slowly. But many functions of interest, such as those required to compute means and variances, are not bounded. So, it is preferable to use stronger, quantitative measures of convergence that also will imply convergence in distribution. To do so, we need to define a notion of distance between probability distribution.

Definition 5.1.1 (Metric)**.** *Given a set $\mathcal{S}$, a function $m : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is called a* **metric** *if the following for all $x, y, z \in \mathcal{S}$:*

1. *$m(x, x) = 0$ (the distance from a point to itself is zero);*
2. *if $x \neq y$, then $m(x, y) > 0$ (the distance between distinct points is positive);*
3. *$m(x, y) = m(y, x)$ (distance is symmetric); and*

4. $m(x, z) \leq m(x, y) + m(y, z)$ *(the triangle inequality holds).*

The pair $(\mathcal{S}, m)$ *is called* **metric space**.

We can use a metric defined on a set a probability distributions to define a new notion of convergence.

Definition 5.1.2 (Metric convergence)**.** *For a metric* M *on set of probability distributions* $\mathcal{M}$*, a sequence of random variables* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ *is said to* **converge in** M *to a random variable* $\boldsymbol{X}_\infty$ *if*

$$\lim_{k \to \infty} M(\mathcal{L}_{\boldsymbol{X}_k}, \mathcal{L}_{\boldsymbol{X}_\infty}) = 0.$$

We then write $\boldsymbol{X}_k \overset{M}{\to} \boldsymbol{X}_\infty$*. Note that we must have* $\mathcal{L}_{\boldsymbol{X}_k} \in \mathcal{M}$ *for* $k \in \mathbb{N} \cup \{\infty\}$ *for these statements to be well-defined.*

We will be particularly interested in the following class of metrics for probability distributions.

Definition 5.1.3. *Fix a metric space* $(\mathcal{A}, m)$ *and a constant* $p \geq 1$*. For the distributions* π *and* π' *on* $\mathcal{A}$*, the* (p, m)**-Wasserstein distance** *is given by*

$$W_{p,m}(\pi, \pi') := \inf_{\gamma} \left\{ \int m(\boldsymbol{x}, \boldsymbol{y})^p \gamma(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \right\}^{\frac{1}{p}},$$

where the infinum is over distributions γ *on* $\mathcal{A} \times \mathcal{A}$ *such that* γ *has marginal distributions* π *and* π'*; that is,* $\pi(\boldsymbol{x}) = \int \gamma(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$ *and* $\pi'(\boldsymbol{y}) = \int \gamma(\mathrm{d}\boldsymbol{x}, \boldsymbol{y})$*.*

Importantly, a coupling γ^* exists that realizes the infinum; that is, such that $W_{p,m}(\pi, \pi') = \{\int m(\boldsymbol{x}, \boldsymbol{y})^p \gamma^*(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\}^{\frac{1}{p}}$. It follows from Jensen's inequality that for $1 \leq p \leq q$, $W_{p,m}(\pi, \pi') \leq W_{q,m}(\pi, \pi')$.

When $p = 1$, there is an alternative characterization of Wasserstein distance that directly relates to the goal of estimation, which is based on the following notion of smoothness.

Definition 5.1.4 (Lipschitz continuity)**.** *Let* m *be a metric on* $\mathcal{A}$*. A function* $\phi : \mathcal{A} \to \mathbb{R}^\ell$ *is* L-Lipschitz (with respect to m) *if* $\|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})\|_2 \leq L\, m(\boldsymbol{x}, \boldsymbol{y})$ *for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$*. We denote the minimal Lipschitz constant* L *by* $\|\phi\|_L$*.*

So, if a function is L-Lipschitz, then it is not changing at more than a linear rate L with respect to the metric m. It turns out that

$$W_{1,m}(\pi, \pi') = \sup_{\substack{\phi:\mathcal{A}\to\mathbb{R} \\ \|\phi\|_L \leq 1}} |\pi(\phi) - \pi'(\phi)|.$$

Thus, if the 1-Wasserstein distance is small, we can accurately approximate the expectation with respect to π of the Lipschitz function ϕ by an expectation with respect to π'.

Two choices of the metric m are of particular interest. The first is $m(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{1}(\boldsymbol{x} \neq \boldsymbol{y})$, which for $p = 1$ induces the *total variation distance* $\mathrm{D}_{\mathrm{TV}}(\pi, \pi')$. In this case ϕ is L-Lipschitz if $\sup_{\boldsymbol{x}} f(\boldsymbol{x}) - \inf_{\boldsymbol{x}} f(\boldsymbol{x}) \leq L$. In other words, total variation distance bounds the difference between expectations of bounded functions.

The second choice is $m(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$, the Euclidean distance, in which case we will use the shorthand notation $W_p(\pi, \pi')$. In this case a differentiable function ϕ is L-Lipschitz if $\|\phi'(\boldsymbol{x})\|_2 \leq L$ for all $\boldsymbol{x}$. Small Euclidean Wasserstein distance can imply small differences in means, covariances, and standard deviations.

Theorem 5.1.5. *For distributions μ and ν on $\mathbb{R}^D$, let $\boldsymbol{X} \sim \mu$ and $\boldsymbol{Y} \sim \nu$. If $W_1(\mu, \nu) < \varepsilon$, then the difference in the means is bounded as*

$$\|\mathbb{E}(\boldsymbol{X}) - \mathbb{E}(\boldsymbol{Y})\|_2 \leq \varepsilon. \tag{5.1}$$

Define the covariance matrices $\boldsymbol{\Sigma} := \mathrm{Cov}(\boldsymbol{X})$ and $\boldsymbol{V} := \mathrm{Cov}(\boldsymbol{Y})$, and let $s := \min\{\|\boldsymbol{\Sigma}\|_2^{1/2}, \|\boldsymbol{V}\|_2^{1/2}\}$. If $W_2(\mu, \nu) < \varepsilon$, then Eq. (5.1) holds, and, furthermore, the difference in marginal standard deviations is bounded as

$$\max_{d\in[D]} |\Sigma_{dd}^{1/2} - V_{dd}^{1/2}| \leq \varepsilon$$

and the difference in the covariances is bounded as

$$\|\boldsymbol{\Sigma} - \boldsymbol{V}\|_2 \leq 2\varepsilon(s + \varepsilon).$$

5.2 Convergence to Stationarity of SGD with Constant Step Size

When the step size is constant, the SGD iterates form a homogenous Markov chain, so it is plausible that the iterates could converge to a stationary

distribution, which we will tentatively denote by ν_η. The error bounds in Chapter 4 for the individual iterates and the iterate average suggest this is true. Let P_η denote the SGD Markov transition kernel with step size η. We now aim to show that the iterates of SGD with a constant step size converge to a unique stationary distribution under weaker assumptions than those required for the error bounds.

Theorem 5.2.1. *If Assumptions 4.5.1 to 4.5.3 hold and $\eta_k = \eta \in (0, 2/L)$, then the SGD iterates converge to a unique stationary distribution ν_η. In addition, for any initial value $\boldsymbol{x}_0$, all $k \geq 0$, and $\tilde{\beta} := 1 - 2\,\eta\mu(1 - \eta L/2)$,*

$$W_2^2(P_\eta^k(\boldsymbol{x}_0, \cdot), \nu_\eta) \leq \tilde{\beta}^k \int \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2^2 \nu_\eta(\mathrm{d}\boldsymbol{x}). \tag{5.2}$$

Proof. The proof makes use of the coupling characterization of the Wasserstein distance. Toward that end, we consider two copies of the SGD iterates, which we denote as $(\boldsymbol{x}_k^{(1)})_{k\in\mathbb{N}}$ and $(\boldsymbol{x}_k^{(2)})_{k\in\mathbb{N}}$. We denote their initial distributions by, respectively, $\nu_0^{(1)}$ and $\nu_0^{(2)}$, and couple the initial iterates so that $W_2^2(\nu_0^{(1)}, \nu_0^{(2)}) = \mathbb{E}(\|\boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}\|_2^2)$. Furthermore, both copies use the same noise:

$$\begin{aligned}
\boldsymbol{x}_{k+1}^{(1)} &\leftarrow \boldsymbol{x}_k^{(1)} - \eta\{f'(\boldsymbol{x}_k^{(1)}) + \varepsilon_{k+1}(\boldsymbol{x}_k^{(1)})\} \\
\boldsymbol{x}_{k+1}^{(2)} &\leftarrow \boldsymbol{x}_k^{(2)} - \eta\{f'(\boldsymbol{x}_k^{(2)}) + \varepsilon_{k+1}(\boldsymbol{x}_k^{(2)})\}.
\end{aligned}$$

By the definition of the Wasserstein distance,

$$\begin{aligned}
W_2^2(\nu_0^{(1)}P_\eta, \nu_0^{(2)}P_\eta) &\leq \mathbb{E}(\|\boldsymbol{x}_1^{(1)} - \boldsymbol{x}_1^{(2)}\|_2^2) \\
&= \mathbb{E}[\|\boldsymbol{x}_0^{(1)} - \eta\widehat{f}_1'(\boldsymbol{x}_0^{(1)}) - \{\boldsymbol{x}_0^{(2)} - \eta\widehat{f}_1'(\boldsymbol{x}_0^{(2)})\}\|_2^2] \\
&= \mathbb{E}[\|\boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}\|_2^2 + \eta^2\|\widehat{f}_1'(\boldsymbol{x}_0^{(1)}) - \widehat{f}_1'(\boldsymbol{x}_0^{(2)})\|_2^2] \\
&\quad - 2\eta\,\mathbb{E}[\langle \boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}, \widehat{f}_1'(\boldsymbol{x}_0^{(1)}) - \widehat{f}_1'(\boldsymbol{x}_0^{(2)})\rangle].
\end{aligned} \tag{5.3}$$

Since the ε_k are independent of $\boldsymbol{x}_k^{(1)}$ and $\boldsymbol{x}_k^{(2)}$, for $i, j \in \{1, 2\}$,

$$\mathbb{E}\{\langle \varepsilon_k(\boldsymbol{x}_k^{(i)}), \boldsymbol{x}_k^{(j)}\rangle\} = \mathbb{E}[\mathbb{E}_k\{\langle \varepsilon_k(\boldsymbol{x}_k^{(i)}), \boldsymbol{x}_k^{(j)}\rangle\}] = 0.$$

Thus, we can rewrite the final expectation in Eq. (5.3) as

$$\begin{aligned}
&\mathbb{E}[\langle \boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}, \widehat{f}_1'(\boldsymbol{x}_0^{(1)}) - \widehat{f}_1'(\boldsymbol{x}_0^{(2)})\rangle] \\
&= \mathbb{E}[\langle \boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}, f'(\boldsymbol{x}_1^{(1)}) + \varepsilon_1(\boldsymbol{x}_0^{(1)}) - f'(\boldsymbol{x}_1^{(2)}) - \varepsilon_1(\boldsymbol{x}_0^{(2)})\rangle] \\
&= \mathbb{E}[\langle \boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}, f'(\boldsymbol{x}_0^{(1)}) - f'(\boldsymbol{x}_0^{(2)})\rangle].
\end{aligned} \tag{5.4}$$

Hence,

$$\begin{aligned} W_2^2(\nu_0^{(1)}P_\eta, \nu_0^{(2)}P_\eta) &\overset{(i)}{\leq} \mathbb{E}[\|\boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}\|_2^2 + \eta^2\|\widehat{f}_1'(\boldsymbol{x}_0^{(1)}) - \widehat{f}_1'(\boldsymbol{x}_0^{(2)})\|_2^2] \\ &\quad - 2\eta\,\mathbb{E}(\langle \boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}, f'(\boldsymbol{x}_0^{(1)}) - f'(\boldsymbol{x}_0^{(2)})\rangle) \\ &\overset{(ii)}{\leq} \mathbb{E}(\|\boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}\|_2^2) \\ &\quad - 2\eta(1 - \eta L/2)\,\mathbb{E}[\langle \boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}, f'(\boldsymbol{x}_0^{(1)}) - f'(\boldsymbol{x}_0^{(2)})\rangle] \\ &\overset{(iii)}{\leq} \{1 - 2\mu\eta(1 - \eta L/2)\}\mathbb{E}(\|\boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}\|_2^2) \\ &= \tilde{\beta}\,\mathbb{E}(\|\boldsymbol{x}_0^{(1)} - \boldsymbol{x}_0^{(2)}\|_2^2), \end{aligned}$$

where (i) follows from Eqs. (5.3) and (5.4), (ii) follows from Assumption 4.5.3, and (iii) follows from Assumption 4.5.2. It follows by an induction that

$$\begin{aligned} W_2^2(\nu_0^{(1)}P_\eta^k, \nu_0^{(2)}P_\eta^k) &\leq \mathbb{E}(\|\boldsymbol{x}_k^{(1)} - \boldsymbol{x}_k^{(2)}\|_2^2) \\ &\leq \tilde{\beta}\,\mathbb{E}(\|\boldsymbol{x}_{k-1}^{(1)} - \boldsymbol{x}_{k-1}^{(2)}\|_2^2) \\ &\leq \tilde{\beta}^k\, W_2^2(\nu_0^{(1)}, \nu_0^{(2)}). \end{aligned} \tag{5.5}$$

We omit the proof that Eq. (5.5) implies that the iterates converge to a unique stationary distribution because it relies on results from real analysis. □

Exercise 5.2.1. *Use Eq. (5.5) and the fact that the the iterates converge to a unique stationary distribution to show Eq. (5.2) holds.*

A benefit of establishing that the SGD iterates converge to a stationary distribution arises from the following characterization of the bias of the stationary distribution mean $\bar{\boldsymbol{x}}_\eta := \int \boldsymbol{x}\nu_\eta(\mathrm{d}\boldsymbol{x})$. Recall that $\varepsilon_k := f' - \widehat{f}_k'$ and let $\boldsymbol{C}(\boldsymbol{x}) := \mathrm{Cov}\{\varepsilon_1(\boldsymbol{x})\}$. We require a new assumption:

Assumption 5.2.2. *The following hold:*

1. *For all $n \in \{2, \ldots, 5\}$, $\sup_{\boldsymbol{x}\in\mathcal{A}}\|f^{(n)}(\boldsymbol{x})\| < \infty$.*
2. *The function $\boldsymbol{C}$ is three-times continuously differentiable and there exist constants $C_\varepsilon, \ell_\varepsilon \geq 0$ such that*

$$\max_{n\in\{1,2,3\}} \|\boldsymbol{C}^{(n)}(\boldsymbol{x})\| \leq C_\varepsilon\{1 + \|\boldsymbol{x} - \boldsymbol{x}_\star\|_2\}^{\ell_\varepsilon}.$$

3. For $p_\varepsilon := \max(6, 2\ell_\varepsilon + 2)$, $\mathbb{E}\{\|\varepsilon_1(\boldsymbol{x}_\star)\|_2^{p_\varepsilon}\} < \infty$.

We write $O(g(\eta))$ to denote a function $r(\eta)$ that, for some $c > 0$, satisfies $\|r(\eta)\|_2 \leq c\, g(\eta)$ for all η sufficiently small.

Theorem 5.2.3. *If Assumptions 4.5.1 to 4.5.3 and 5.2.2 hold, then there exists a vector* $\boldsymbol{v}$ *such that*

$$\bar{\boldsymbol{x}}_\eta - \boldsymbol{x}_\star = \eta \boldsymbol{v} + O(\eta^2).$$

We will not prove this particular result. Instead, we focus on an important implication when combined with Theorem 5.2.1, which is that we can dramatically improve accuracy by using ***Richardson–Romberg extrapolation***. The idea is to combine estimates of $\bar{\boldsymbol{x}}_\eta$ and $\bar{\boldsymbol{x}}_{2\,\eta}$ such that that the $O(\eta)$ errors cancel out, resulting in an estimate of $\boldsymbol{x}_\star$ with $O(\eta^2)$ error. First note that it follows from Theorems 5.1.5 and 5.2.1 that for $B(\eta, \boldsymbol{x}_0) := \int \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2^2 \nu_\eta(\mathrm{d}\boldsymbol{x})$,

$$|\mathbb{E}(\boldsymbol{x}_k^{(\eta)}) - \bar{\boldsymbol{x}}_\eta| \leq B(\eta, \boldsymbol{x}_0)\tilde{\beta}^{k/2}.$$

Therefore,

$$\begin{aligned}
|\mathbb{E}(\bar{\boldsymbol{x}}_{0:k}^{(\eta)}) - \bar{\boldsymbol{x}}_\eta| &\leq \frac{1}{k}\sum_{\ell=0}^{k-1} |\mathbb{E}(\boldsymbol{x}_\ell^{(\eta)}) - \bar{\boldsymbol{x}}_\eta| \\
&\leq \frac{1}{k} B(\eta, \boldsymbol{x}_0) \sum_{\ell=0}^{k-1} \tilde{\beta}^{\ell/2} \\
&\leq \frac{1}{k} \underbrace{\frac{B(\eta, \boldsymbol{x}_0)}{1 - \tilde{\beta}^{1/2}}}_{C(\eta, \boldsymbol{x}_0)}.
\end{aligned}$$

Combining this bound with Theorem 5.2.3, we have

$$\mathbb{E}(\bar{\boldsymbol{x}}_{0:k}^{(\eta)}) - \boldsymbol{x}_\star = \eta \boldsymbol{A} + r(\eta, \boldsymbol{x}_0, k),$$

where $r(\eta, \boldsymbol{x}_0, k)$ is a remainder term bounded as

$$\|r(\eta, \boldsymbol{x}_0, k)\|_2 \leq c\,\eta^2 + C(\eta, \boldsymbol{x}_0)/k.$$

Therefore, we can use the extrapolated estimate $2\,\bar{\boldsymbol{x}}_{0:k}^{(\eta)} - \bar{\boldsymbol{x}}_{0:k}^{(2\,\eta)}$ of $\boldsymbol{x}_\star$, which has bias bounded as

$$\begin{aligned}
\|\mathbb{E}\{2\bar{\boldsymbol{x}}_{0:k}^{(\eta)} - \bar{\boldsymbol{x}}_{0:k}^{(2\,\eta)} - \boldsymbol{x}_\star\}\|_2 &= \|2\, r(\eta, \boldsymbol{x}_0, k) - r(2\,\eta, \boldsymbol{x}_0, k)\|_2 \\
&\leq 6\, c\, \eta^2 + \frac{2C(\eta, \boldsymbol{x}_0) + C(2\,\eta, \boldsymbol{x}_0)}{k}. \qquad (5.6)
\end{aligned}$$

Letting $\delta_{(k)} := \tilde{\beta}^{k/2}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2$, we thus have the following summary of the expected error of three estimators for $\boldsymbol{x}_\star$ using constant-step size SGD, which only reports dependence on η and k:

Estimator	$\boldsymbol{x}_k$	$\bar{\boldsymbol{x}}_{k/2:k}$	$2\bar{\boldsymbol{x}}^{(\eta/2)}_{k/2:k} - \bar{\boldsymbol{x}}^{(\eta)}_{k/2:k}$
Error bound	Eq. (4.8)	Eq. (4.9)	Eq. (5.6)
Irreducible error	$O(\eta^{1/2})$	$O(\eta)$	$O(\eta^2)$
Initialization-dependent error	$\delta_{(k)}$	$O\left(\frac{\delta_{(k)}}{\eta k}\right)$	–
Other transient error	0	$O\left(\frac{1}{k^{1/2}} + \frac{1}{\eta k}\right)$	–

The transient errors terms for the Richardson–Romberg extrapolation are omitted, as we have not explicitly developed them. But we should expect them to be of the same order as the in the iterate average case, since it is formed by taking the difference of iterate averages.

Chapter 6

Designing Markov Chains

With Applications to Developing MCMC Algorithms

Develops tools for defining Markov chains with a desired stationary distribution including detailed balance and reversibility, and compositions and mixtures of kernels. As an application, introduced two classes of MCMC algorithms: the Metropolis–Hastings algorithm and the Gibbs sampler.

6.1 Detailed Balance

Definition 6.1.1 (p.d.f. and p.m.f. notation)**.** *If a distribution (e.g., μ) has a p.d.f. (or p.m.f.), we will denote as h_μ. We will always denote probability kernels using uppercase letters (e.g., P). If for all $\boldsymbol{x} \in \mathcal{A}$ the kernel (e.g., $P(\boldsymbol{x}, \cdot)$) has a p.d.f. (or p.m.f.), we will denote it using the a lowercase letter (e.g., $p(\boldsymbol{x}, \boldsymbol{y})$). To streamline discussions, we will refer to both p.d.f.s and p.m.f.s as "densities," which the understanding that we will not mix p.d.f.s and p.m.f.s unless stated explicitly.*

Checking the stationarity condition $\nu = \nu P$ directly can be cumbersome because it involves a (potentially complicated) integral. A particularly convenient sufficient condition for stationarity is called detailed balance. For distribution μ, define the ***product probability measure*** $\mu \otimes P$ on $\mathcal{A} \times \mathcal{A}$ such that for $A, B \in \mathcal{P}(\mathcal{A})$,

$$(\mu \otimes P)(A \times B) = \int_A \int_B P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \mu(\mathrm{d}\boldsymbol{x}).$$

In particular, if $\boldsymbol{X} \sim \mu$ and $\boldsymbol{Y}$ has conditional distribution $\boldsymbol{Y} \mid \boldsymbol{X} \sim P(\boldsymbol{X}, \cdot)$,

then $(\boldsymbol{X}, \boldsymbol{Y})$ has joint distribution $\mu \otimes P$. If μ and P both have densities, then density of $\mu \otimes P$ is $\mu(\boldsymbol{x})P(\boldsymbol{x}, \boldsymbol{y})$.

Definition 6.1.2 (Detailed balance)**.** *The Markov chain with transition kernel P is said to satisfy* **detailed balance with respect to** *ν if, for all $A, B \in \mathcal{P}(\mathcal{A})$,*

$$(\nu \otimes P)(A \times B) = (\nu \otimes P)(B \times A).$$

Lemma 6.1.3. *When densities exist, Definition 6.1.2 is equivalent to: for some $S \in \mathcal{P}(\mathcal{A})$ with $\nu(S) = 1$ and for all $\boldsymbol{x}, \boldsymbol{y} \in S$,*

$$h_\nu(\boldsymbol{x})p(\boldsymbol{x}, \boldsymbol{y}) = h_\nu(\boldsymbol{y})p(\boldsymbol{y}, \boldsymbol{x}).$$

Proposition 6.1.4. *If P satisfies detailed balance with respect to ν, then ν is an invariant distribution of the Markov chain with kernel P.*

Proof. Using the definition of νP and the fact that by construction $P(\boldsymbol{x}, B) = \int_B P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$, we have

$$\begin{aligned}\nu P(B) &= \int_{\mathcal{A}} P(\boldsymbol{x}, B)\nu(\mathrm{d}\boldsymbol{x}) \\ &= \int_{\mathcal{A}} \int_B P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\nu(\mathrm{d}\boldsymbol{x}).\end{aligned}$$

Then using the definition of $\nu \otimes P$, the detailed balance assumption, and the fact that $P(\boldsymbol{x}, \cdot)$ and ν are probability measures, we have

$$\begin{aligned}\int_{\mathcal{A}} \int_B P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\nu(\mathrm{d}\boldsymbol{x}) &= (\nu \otimes P)(\mathcal{A} \times B) \\ &= (\nu \otimes P)(B \times \mathcal{A}) \\ &= \int_B \int_{\mathcal{A}} P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\nu(\mathrm{d}\boldsymbol{x}) \\ &= \int_B \nu(\mathrm{d}\boldsymbol{x}) \\ &= \nu(B).\end{aligned}$$

□

If a Markov chain satisfies detailed balance, it is said to be ***reversible with respect to*** ν: when $\boldsymbol{X}_0 \sim \nu$, for all $k \in \mathbb{N}$, $(\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}, \boldsymbol{X}_k) \overset{d}{=} (\boldsymbol{X}_k, \boldsymbol{X}_{k-1}, \ldots, \boldsymbol{X}_1, \boldsymbol{X}_0)$. In other words, the distribution of the Markov chain remains the same if it is "run in reverse."

Example 6.1.5 (Detailed balance for a finite-state Markov chain)**.** *Following the set-up and notation from Examples 3.2.2 and 3.3.2, we can write the detailed balance condition as requiring that, for all $d, d' \in [D]$,*

$$\pi_d K_{d,d'} = \pi_{d'} K_{d',d}.$$

As a first example, consider a Markov chain that stays at its current state, goes to the previous state, or goes to the next state with equal probability. More formally, the Markov chain has the transition kernel given by

$$K_{d,d'} = \begin{cases} 1/3 & \text{if } d' = d - 1 \mod D \\ 1/3 & \text{if } d' = d \\ 1/3 & \text{if } d' = d \mod D + 1 \\ 0 & \text{otherwise.} \end{cases}$$

This Markov chain's unique invariant distribution is the uniform distribution $\pi_d = 1/D$. It satisfies detailed balance as well. For example, for $D \geq 3$,

$$\pi_1 K_{1,2} = \frac{1}{D} \times \frac{1}{3} = \pi_2 K_{2,1}$$

and

$$\pi_1 K_{1,3} = \frac{1}{D} \times 0 = \pi_3 K_{3,1}.$$

As a second example, consider a Markov chain that stays at its current state or goes to the previous state with equal probability. More formally, the Markov chain has the transition kernel given by

$$K_{d,d'} = \begin{cases} 1/2 & \text{if } d' = d \\ 1/2 & \text{if } d' = d \mod D + 1 \\ 0 & \text{otherwise.} \end{cases}$$

This Markov chain's unique invariant distribution is the uniform distribution but it does note satisfy detailed balance since it can only move in one direction. For example, for $D \geq 2$,

$$\pi_1 K_{1,2} = \frac{1}{D} \times \frac{1}{2}$$

but

$$\pi_2 K_{2,1} = \frac{1}{D} \times 0 = 0.$$

Example 6.1.6 (Gaussian AR(1) process, continued)**.** *The AR(1) process from Example 3.3.3 is reversible with respect to the invariant distribution* $\nu = \mathcal{N}(\mathbf{0}, \{1-\alpha^2\}^{-1}\boldsymbol{\Sigma}_\varepsilon)$*. To see since, we calculate*

$$
\begin{aligned}
&h_\nu(\boldsymbol{x})p(\boldsymbol{x},\boldsymbol{y}) \\
&= \log \mathcal{N}\big(\boldsymbol{x} \,\big|\, 0, \{1-\alpha^2\}^{-1}\boldsymbol{\Sigma}_\varepsilon\big)\mathcal{N}(\boldsymbol{y} \mid \alpha\boldsymbol{x}, \boldsymbol{\Sigma}_\varepsilon^2) \\
&= -\frac{1-\alpha^2}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{y}-\alpha\boldsymbol{x})^\top\boldsymbol{\Sigma}_\varepsilon^{-1}(\boldsymbol{y}-\alpha\boldsymbol{x}) + \text{constant} \\
&= -\frac{1}{2}\Big\{(1-\alpha^2)\boldsymbol{x}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{x} + \boldsymbol{y}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{y} - 2\alpha\boldsymbol{x}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{y} + \alpha^2\boldsymbol{x}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{x})\Big\} + \text{constant} \\
&= -\frac{1}{2}\Big\{\boldsymbol{x}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{x} + \boldsymbol{y}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{y} - 2\alpha\boldsymbol{x}^\top\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{y}\Big\} + \text{constant}.
\end{aligned}
$$

Since the final expression is symmetric in $\boldsymbol{x}$ *and* $\boldsymbol{y}$*, we conclude that* $h_\nu(\boldsymbol{x})p(\boldsymbol{x},\boldsymbol{y}) = h_\nu(\boldsymbol{y})p(\boldsymbol{y},\boldsymbol{x})$*.*

6.2 Combining Markov Kernels

For MCMC, we will often want to combine Markov kernels that have the same invariant distribution. We previously saw how to combine kernels via composition. Another way to combine kernels is via linear combination.

Definition 6.2.1. *For* $w_1, w_2 \geq 0$*, the* **mixture** $w_1P_1 + w_2P_2$ *of kernels* P_1 *and* P_2 *is defined as*

$$(w_1P_1 + w_2P_2)(\boldsymbol{x}, A) := w_1P_1(\boldsymbol{x}, A) + w_2P_2(\boldsymbol{x}, A).$$

If $w_1 + w_2 = 1$, then mixing can be interpreted as transitioning according to P_1 with probability w_1 and according to P_2 otherwise (with probability $w_2 = 1 - w_1$).

Lemma 6.2.2. *If* P_1 *and* P_2 *are valid transition kernels, then so is* $wP_1 + (1-w)P_2$ *with* $w \in [0,1]$*.*

Exercise 6.2.1. *Prove Lemma 6.2.2*

Like composition, mixing is associative: $(w_1P_1 + w_2P_2) + w_3P_3 = w_1P_1 + (w_2P_2 + w_3P_3)$. So, we can write $w_1P_1 + w_2P_2 + w_3P_3$ without any ambiguity.

Composition and mixing preserve invariant distributions.

Lemma 6.2.3. *If P_1 and P_2 have invariant distribution ν, then so do P_1P_2 and $wP_1 + (1-w)P_2$ (for $w \in (0,1)$).*

Exercise 6.2.2 (Composition and mixing preserve invariant distributions)**.** *Prove Lemma 6.2.3.*

However, while mixing preserves reversibility, composition does not. This makes some intuitive sense since transitioning according to P_1 then P_2 is not the same as (when running the chain backward) transitioning according to P_2 then P_1. However, the composition $P_1P_2P_1$ does preserve reversibility.

Lemma 6.2.4. *If P_1 and P_2 are reversible with respect to ν, then so are $P_1P_2P_1$ and $wP_1 + (1-w)P_2$ (for $w \in (0,1)$).*

Proof. Using the definition of composition and the reversibility assumption,

$$
\begin{aligned}
&\nu \otimes P_1P_2P_1(A \times B) \\
&= \int \mathbb{1}(\boldsymbol{x} \in A, \boldsymbol{y}' \in \mathcal{A}, \boldsymbol{y}'' \in \mathcal{A}, \boldsymbol{y} \in B)\nu(\mathrm{d}\boldsymbol{x})P_1(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}')P_2(\boldsymbol{y}', \mathrm{d}\boldsymbol{y}'')P_1(\boldsymbol{y}'', \mathrm{d}\boldsymbol{y}) \\
&= \int \mathbb{1}(\boldsymbol{x} \in A, \boldsymbol{y}' \in \mathcal{A}, \boldsymbol{y}'' \in \mathcal{A}, \boldsymbol{y} \in B)\nu(\mathrm{d}\boldsymbol{y}')P_1(\boldsymbol{y}', \mathrm{d}\boldsymbol{x})P_2(\boldsymbol{y}', \mathrm{d}\boldsymbol{y}'')P_1(\boldsymbol{y}'', \mathrm{d}\boldsymbol{y}) \\
&= \int \mathbb{1}(\boldsymbol{x} \in A, \boldsymbol{y}' \in \mathcal{A}, \boldsymbol{y}'' \in \mathcal{A}, \boldsymbol{y} \in B)\nu(\mathrm{d}\boldsymbol{y}'')P_2(\boldsymbol{y}'', \mathrm{d}\boldsymbol{y}')P_1(\boldsymbol{y}', \mathrm{d}\boldsymbol{x})P_1(\boldsymbol{y}'', \mathrm{d}\boldsymbol{y}) \\
&= \int \mathbb{1}(\boldsymbol{x} \in A, \boldsymbol{y}' \in \mathcal{A}, \boldsymbol{y}'' \in \mathcal{A}, \boldsymbol{y} \in B)\nu(\mathrm{d}\boldsymbol{y})P_1(\boldsymbol{y}, \mathrm{d}\boldsymbol{y}'')P_2(\boldsymbol{y}'', \mathrm{d}\boldsymbol{y}')P_1(\boldsymbol{y}', \mathrm{d}\boldsymbol{x}) \\
&= \nu \otimes P_1P_2P_1(B \times A).
\end{aligned}
$$

□

Exercise 6.2.3. *Prove Lemma 6.2.4 for $wP_1 + (1-w)P_2$.*

6.3 Markov Chain Monte Carlo

The most widely used MCMC method is the Metropolis–Hastings (MH) algorithm. Construction of the MH transition kernel P depends on the choice of a base Markov kernel $Q : \mathcal{A} \times \mathcal{P}(\mathcal{A}) \to [0,1]$ called the ***proposal***

distribution. Assume that π and $Q(\boldsymbol{x}, \cdot)$ have densities. Given the current state $\boldsymbol{X}_k$, procedurally the first step is to sample a ***proposal*** $\boldsymbol{X}'_{k+1} \sim Q(\boldsymbol{X}_k, \cdot)$. Define the ***Metropolis–Hastings acceptance probability***

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \min\left\{\frac{h_\pi(\boldsymbol{y})q(\boldsymbol{y}, \boldsymbol{x})}{h_\pi(\boldsymbol{x})q(\boldsymbol{x}, \boldsymbol{y})}, 1\right\}.$$

With probability $\alpha(\boldsymbol{X}_k, \boldsymbol{X}'_{k+1})$ set $\boldsymbol{X}_{k+1} = \boldsymbol{X}'_{k+1}$ (the proposal is ***accepted***) and otherwise set $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k$ (the proposal is ***rejected***).

Intuitively, a proposal is more likely to be accepted when the new state has higher probability and the probability of transitioning from the new state to the old state is large compared to the reverse transition. Since $\alpha(\boldsymbol{x}, \boldsymbol{y})$ only depends on π through the ratio $h_\pi(\boldsymbol{y})/h_\pi(\boldsymbol{x})$, the p.d.f. (or p.m.f.) h_π needs to be known only up to a multiplicative constant.

The MH Markov transition kernel is given by

$$P(\boldsymbol{x}, A) = \int_A \alpha(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{x}, \boldsymbol{y})\mu(\mathrm{d}\boldsymbol{y}) + \bar{\alpha}(\boldsymbol{x})\delta_{\boldsymbol{x}}(A), \tag{6.1}$$

where

$$\bar{\alpha}(\boldsymbol{x}) = \int_{\mathcal{A}} \{1 - \alpha(\boldsymbol{x}, \boldsymbol{y})\}q(\boldsymbol{x}, \boldsymbol{y})\mu(\mathrm{d}\boldsymbol{y})$$

is the probability of rejection. The kernel satisfies detailed balance with respect to π, and hence is reversible and has invariant distribution π.

Proposition 6.3.1 (The MH kernel satisfies detailed balance)**.** *The MH kernel defined in Eq.* (6.1) *satisfies detailed balance with respect to* π*.*

Proof. We have

$$\begin{aligned}
&\pi \otimes P(A \times B) \\
&= \underbrace{\int_A \int_B h_\pi(\boldsymbol{x})\alpha(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{x}, \boldsymbol{y})\mu(\mathrm{d}\boldsymbol{y})\mu(\mathrm{d}\boldsymbol{x})}_{\star} + \underbrace{\int_A \int_B h_\pi(\boldsymbol{x})\bar{\alpha}(\boldsymbol{x})\delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y})\mu(\mathrm{d}\boldsymbol{x})}_{\star\star}.
\end{aligned}$$

Observe that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$,

$$\begin{aligned}
h_\pi(\boldsymbol{x})\alpha(\boldsymbol{x}, \boldsymbol{y})q(\boldsymbol{x}, \boldsymbol{y}) &= h_\pi(\boldsymbol{x})q(\boldsymbol{x}, \boldsymbol{y}) \min\left\{\frac{h_\pi(\boldsymbol{y})q(\boldsymbol{y}, \boldsymbol{x})}{h_\pi(\boldsymbol{x})q(\boldsymbol{x}, \boldsymbol{y})}, 1\right\} \\
&= \min\{h_\pi(\boldsymbol{y})q(\boldsymbol{y}, \boldsymbol{x}), h_\pi(\boldsymbol{x})q(\boldsymbol{x}, \boldsymbol{y})\} \\
&= h_\pi(\boldsymbol{y})\alpha(\boldsymbol{y}, \boldsymbol{x})q(\boldsymbol{y}, \boldsymbol{x}),
\end{aligned}$$

so

$$\begin{aligned}\star &= \int_A \int_B h_\pi(\boldsymbol{y})\alpha(\boldsymbol{y},\boldsymbol{x})q(\boldsymbol{y},\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{y})\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_B \int_A h_\pi(\boldsymbol{y})\alpha(\boldsymbol{y},\boldsymbol{x})q(\boldsymbol{y},\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{y}).\end{aligned}$$

On the other hand,

$$\begin{aligned}\star\star &= \int_A h_\pi(\boldsymbol{x})\bar{\alpha}(\boldsymbol{x})\mathbb{1}(\boldsymbol{x}\in B)\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int h_\pi(\boldsymbol{x})\bar{\alpha}(\boldsymbol{x})\mathbb{1}(\boldsymbol{x}\in A\cap B)\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_B h_\pi(\boldsymbol{x})\bar{\alpha}(\boldsymbol{x})\mathbb{1}(\boldsymbol{x}\in A)\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_B \int_A h_\pi(\boldsymbol{x})\bar{\alpha}(\boldsymbol{x})\delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y})\mu(\mathrm{d}\boldsymbol{x}).\end{aligned}$$

Combining the two previous displays, conclude that $\pi \otimes P(A \times B) = \pi \otimes P(B \times A)$. □

For now we restrict ourselves to three canonical choices for the proposal distribution.

Example 6.3.2 (Independent Metropolis–Hasting algorithm)**.** *Given a distribution Q_0, the independent MH algorithm uses proposal $Q(\boldsymbol{x},\cdot) = Q_0$ that is independent of the current state $\boldsymbol{x}$.*

Example 6.3.3 (Random-walk Metropolis–Hasting algorithm)**.** *Let q_0 be the density of a mean-zero, symmetric random variable such as $\mathcal{N}(0,\sigma^2 I)$. The random-walk MH algorithm uses proposal with density $q(\boldsymbol{x},\boldsymbol{y}) = q_0(\boldsymbol{x}-\boldsymbol{y})$. Hence, before the MH correction, the proposal behaves like a random walk.*

Example 6.3.4 ((Two-stage) Gibbs sampler)**.** *The two-stage Gibbs sampler is based on breaking up the parameter into two pieces, $\boldsymbol{x} = (\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)})$, such that the conditional densities $h_\pi(\boldsymbol{x}_{(1)} \mid \boldsymbol{x}_{(2)})$ and $h_\pi(\boldsymbol{x}_{(2)} \mid \boldsymbol{x}_{(1)})$ are easy to sample from. First, consider using the proposal density $q_1(\boldsymbol{x},\boldsymbol{y}) = h_\pi(\boldsymbol{y}_{(1)} \mid$*

$\boldsymbol{x}_{(2)})\mathbb{1}(\boldsymbol{y}_{(2)} = \boldsymbol{x}_{(2)})$. *In this case the acceptance probability is always 1 since*

$$\begin{aligned}\frac{h_\pi(\boldsymbol{y})q_1(\boldsymbol{y},\boldsymbol{x})}{h_\pi(\boldsymbol{x})q_1(\boldsymbol{x},\boldsymbol{y})} &= \frac{h_\pi(\boldsymbol{y})h_\pi(\boldsymbol{x}_{(1)} \mid \boldsymbol{y}_{(2)})\mathbb{1}(\boldsymbol{y}_{(2)} = \boldsymbol{x}_{(2)})}{h_\pi(\boldsymbol{x})h_\pi(\boldsymbol{y}_{(1)} \mid \boldsymbol{x}_{(2)})\mathbb{1}(\boldsymbol{y}_{(2)} = \boldsymbol{x}_{(2)})} \\ &= \frac{h_\pi(\boldsymbol{y}_{(1)} \mid \boldsymbol{y}_{(2)})h_\pi(\boldsymbol{y}_{(2)})h_\pi(\boldsymbol{x}_{(1)} \mid \boldsymbol{y}_{(2)})\mathbb{1}(\boldsymbol{y}_{(2)} = \boldsymbol{x}_{(2)})}{h_\pi(\boldsymbol{x}_{(1)} \mid \boldsymbol{x}_{(2)})h_\pi(\boldsymbol{x}_{(2)})h_\pi(\boldsymbol{y}_{(1)} \mid \boldsymbol{x}_{(2)})\mathbb{1}(\boldsymbol{y}_{(2)} = \boldsymbol{x}_{(2)})} \\ &= \frac{h_\pi(\boldsymbol{y}_{(1)} \mid \boldsymbol{x}_{(2)})h_\pi(\boldsymbol{x}_{(2)})h_\pi(\boldsymbol{x}_{(1)} \mid \boldsymbol{x}_{(2)})}{h_\pi(\boldsymbol{x}_{(1)} \mid \boldsymbol{x}_{(2)})h_\pi(\boldsymbol{x}_{(2)})h_\pi(\boldsymbol{y}_{(1)} \mid \boldsymbol{x}_{(2)})} \\ &= 1.\end{aligned}$$

Similarly, using the proposal $q_2(\boldsymbol{x},\boldsymbol{y}) = h_\pi(\boldsymbol{y}_{(2)} \mid \boldsymbol{x}_{(1)})\mathbb{1}(\boldsymbol{y}_{(1)} = \boldsymbol{x}_{(1)})$ *also has acceptance probability 1. Thus,* Q_1 *and* Q_2 *are reversible transition kernels with invariant distribution* π*. Hence, by Lemma 6.2.3, the* **two-stage Gibbs transition kernel** $P_G = Q_1Q_2$ *also has invariant distribution* π*. However,* P_G *is not reversible. On the other hand, by Lemma 6.2.4, the randomized Gibbs kernel* $P_{RG} = 0.5Q_1 + 0.5Q_2$ *and the symmetrized Gibbs kernel* $P_{SG} = Q_1Q_2Q_1$ *are reversible.*

Example 6.3.5 (Gibbs sampler for a normal-gamma model)**.** *Consider a model for observed data* $Y \in \mathbb{R}^N$ *with parameters* $\boldsymbol{x} = (m, \tau)$*, where* $m \in \mathbb{R}$ *is the mean and* $\tau \in \mathbb{R}_+$ *is the precision (inverse variance) of a normal distribution:*

$$\begin{aligned}\tau &\sim \mathrm{Gam}(a,b) \\ m \mid \tau &\sim \mathcal{N}(0, 1/(\lambda\tau)) \\ Y_n \mid m,\tau &\sim \mathcal{N}(m, 1/\tau) \qquad\qquad (n = 1,\dots,N),\end{aligned}$$

where the density of the gamma distribution $\mathrm{Gam}(a,b)$ *is* $\mathrm{Gam}(t \mid a,b) = b^a t^{a-1}e^{-bt}/\Gamma(a)$ *(with* $t \in \mathbb{R}_+$*). The hyperparameters* λ*,* a*, and* b *are considered fixed constants. To derive the Gibbs sampler for this model, we must find the conditional distributions* $\tau \mid m, Y$ *and* $m \mid \tau, Y$*. First, we write out the log joint density:*

$$\begin{aligned}\log p(\tau, m, Y) = \underbrace{(a-1)\log\tau - b\tau}_{\text{from } \mathrm{Gam}(\tau|a,b)} + \underbrace{-\frac{\lambda\tau}{2}m^2 + \frac{1}{2}\log\tau}_{\text{from } \mathcal{N}(m|0,1/(\lambda\tau))} \\ + \underbrace{-\frac{\tau}{2}\sum_{n=1}^{N}(Y_n - m)^2 + \frac{N}{2}\log\tau}_{\text{from } \mathcal{N}(Y_n|m,1/\tau)} + c,\end{aligned}$$

where c is a constant that depends only on the (fixed) model hyperparameters λ, a, and b. Now, to find the conditional distribution $\tau \mid m, Y$, we need only consider the terms that involve τ, which after gathering terms with the same τ dependence, yields

$$\log p(\tau \mid m, Y) = \left(a + \frac{N+1}{2} - 1\right) \log \tau - \left\{b + \frac{1}{2} \sum_{n=1}^{N} (Y_n - m)^2\right\} \tau + c',$$

where c' is a constant the does not depend on τ. Recognizing this as the log density of a gamma distribution, we conclude that $\tau \mid m, Y \sim \mathrm{Gam}(a_N, b_N)$, where $a_N := a + \frac{N+1}{2}$ and $b_N := b + \frac{1}{2} \sum_{n=1}^{N} (Y_n - m)^2$. Similarly, for the conditional distribution $m \mid \tau, Y$, after some algebraic manipulations, yields

$$\log p(m \mid \tau, Y) = -\frac{\tau(N+\lambda)}{2} \left(m - \frac{1}{N+\lambda} \sum_{n=1}^{N} Y_i\right)^2 + c'',$$

where c'' is a constant the does not depend on m. Recognizing this as the log density of a normal distribution, we conclude that $m \mid \tau, Y \sim \mathcal{N}(\bar{Y}_N, v_N)$, where $\bar{Y}_N := \frac{1}{N+\lambda} \sum_{n=1}^{N} Y_i$ and $v_N := \frac{1}{\tau(N+\lambda)}$.

Exercise 6.3.1 (Gibbs sampler for a nonnegative matrix factorization model)**.** *Consider the a nonnegative matrix factorization model for observed count data $Y \in \mathbb{N}^{N \times D}$ and parameters $\boldsymbol{x} = (L, R, Z)$, where $L \in \mathbb{R}_+^{N \times K}$, $R \in [0,1]^{K \times D}$ $\sum_{d=1}^{D} R_{kd} = 1$, and $Z \in \mathbb{N}^{N \times D \times K}$. The idea is that the data are explained as linear combinations of K latent factors $R_k = (R_{k1}, \ldots, R_{kD})$ $(k = 1, \ldots, K)$. The parameter L_{nk} represents the loading of the kth factor on the nth observation $Y_n = (Y_{n1}, \ldots, Y_{nD})$.*

$$\begin{aligned}
L_{nk} &\overset{ind}{\sim} \mathrm{Gam}(a, b) && (n = 1, \ldots, N; k = 1, \ldots, K) \\
R_k &\overset{ind}{\sim} \mathrm{Dir}(\alpha, \ldots, \alpha) && (k = 1, \ldots, K) \\
Y_{nd} \mid L, R &\overset{ind}{\sim} \mathrm{Poiss}\left(\textstyle\sum_{k=1}^{K} L_{nk} R_{kd}\right) && (n = 1, \ldots, N; d = 1, \ldots, D) \\
Z_{nd} \mid L, R, Y &\overset{ind}{\sim} \mathrm{Multi}(Y_{nd}, P_{nd}) && (n = 1, \ldots, N; d = 1, \ldots, D),
\end{aligned}$$

where $\overset{ind}{\sim}$ denotes distributed independently of other random variables and $P_{nd} = (L_{nk} R_{kd} / (\sum_{\ell=1}^{K} L_{n\ell} R_{\ell d}))_{k=1}^{K}$.[a] *The hyperparameters a, b, and α are considered fixed constants. The conditional distribution $Z_{nd} \mid L, R, Y$ is given. Compute the other two conditional distributions required for the*

Gibbs sampler, namely the distributions of $L_{nk} \mid R, Z, Y$ and $R_k \mid L, Z, Y$.

[a]For $\alpha \in \mathbb{R}_+^D$ and $t \in \mathbb{R}_+^D$ with $\sum_{d=1}^D t_d = 1$, the density of the Dirichlet distribution $\mathrm{Dir}(\alpha_1, \ldots, \alpha_D)$ is given by $\mathrm{Dir}(t_1, \ldots, t_D \mid \alpha_1, \ldots, \alpha_D) = \Gamma(\sum_{d=1}^D \alpha_d) \prod_{d=1}^D \frac{t_d^{\alpha_d - 1}}{\Gamma(\alpha_d)}$. For $\lambda \in \mathbb{R}_+$ and $y \in \mathbb{N}$, the p.m.f. of the Poisson distribution $\mathrm{Poiss}(\lambda)$ is given by $\mathrm{Poiss}(y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$. For $p \in \mathbb{R}_+^K$ with $\sum_{k=1}^K p_k = 1$, $M \in \mathbb{N}$, and $z \in \mathbb{N}^K$ with $\sum_{k=1}^K z_k = M$ the p.m.f. of the multinomial distribution $\mathrm{Multi}(M, (p_1, \ldots, p_K))$ is given by $\mathrm{Multi}(z \mid M, (p_1, \ldots, p_K)) = \binom{M}{z} \prod_{k=1}^K p_k^{z_k}$.

Chapter 7

Small Sets and Irreducibility

With Applications to Markov Chain Law of Large Numbers

7.1 Irreducibility and the Law of Large Numbers

We begin by investigating when a Markov chain satisfies a strong law of large numbers (LLN) similar to the i.i.d. version given in Theorem 2.6.11. For a Markov chain to satisfy an LLN, it must be ***ergodic***, which, informally, means that the Markov will not get "stuck" in some part of the state space and thus fail to visit "the whole state space." The transition kernel $P(\boldsymbol{x}, A) = \delta_{\boldsymbol{x}}(A)$ from Exercise 3.3.3 produces a non-ergodic Markov chain since $\mathbb{P}(\boldsymbol{X}_k = \boldsymbol{x} \mid \boldsymbol{X}_0 = \boldsymbol{x}) = 1$: the chain is almost surely constant. To avoid such pathologies, we require an ***irreducibility*** assumption, which guarantees there is a specific kind of set, called a ***small set***, which the Markov chain will always return to.

Definition 7.1.1 (Small set)**.** *Call $C \in \mathcal{P}(\mathcal{A})$ a* **small set** *with respect to a probability measure μ if there exists $\beta \in (0, 1)$ such that, for all $\boldsymbol{x} \in C$ and $A \in \mathcal{P}(\mathcal{A})$,*

$$P(\boldsymbol{x}, A) \geq \beta\mu(A).$$

The key feature of a small set is that, when the chain is in the small set, it transitions "somewhat uniformly" like the probability measure μ scaled down by β. For an LLN to hold, we must ensure that, no matter where the Markov chain, it will eventually reach a small set.

Assumption 7.1.2 (Irreducibility)**.** *There exists a small set C such that for all $\boldsymbol{x} \in \mathcal{A}$, there exists $k(\boldsymbol{x}) \in \mathbb{N}$ such that*

$$P^{k(\boldsymbol{x})}(\boldsymbol{x}, C) > 0.$$

Example 7.1.3 (Irreducibility of an AR(1) process)**.** *Recall the AR(1) process from Example 3.1.5 given by*

$$\boldsymbol{X}_k = \alpha \boldsymbol{X}_{k-1} + \varepsilon_k,$$

where $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_k)_{k\in\mathbb{N}}$ is a i.i.d. sequence of random variables satisfying $\mathbb{E}(\boldsymbol{\varepsilon}_0) = 0$. Assume that $\boldsymbol{\varepsilon}_k$ has a continuous, everywhere positive p.d.f. f. For $A \in \mathcal{P}(\mathbb{R}^D)$, define the volume function[1]

$$\lambda(A) := \int_A \mathrm{d}\boldsymbol{x} := \int \mathbb{1}(\boldsymbol{x} \in A)\mathrm{d}\boldsymbol{x}.$$

For some $r > 0$, let $C = \{\boldsymbol{x} \in \mathbb{R}^D : \|\boldsymbol{x}\|_2 \leq r\}$. Then for all $\boldsymbol{x} \in C$ and $A \in \mathcal{P}(\mathbb{R}^D)$,

$$\begin{aligned} P(\boldsymbol{x}, A) &= \int_A f(\boldsymbol{y} - \alpha\boldsymbol{x})\mathrm{d}\boldsymbol{y} \\ &\geq \int_{A\cap C} f(\boldsymbol{y} - \alpha\boldsymbol{x})\mathrm{d}\boldsymbol{y} \\ &\geq \inf_{\boldsymbol{y},\boldsymbol{y}'\in C} f(\boldsymbol{y} - \alpha\boldsymbol{y}') \times \int_{A\cap C} \mathrm{d}\boldsymbol{y} \\ &\geq \lambda(C) \inf_{\boldsymbol{y},\boldsymbol{y}'\in C} f(\boldsymbol{y} - \alpha\boldsymbol{y}') \times \frac{\lambda(A\cap C)}{\lambda(C)}. \end{aligned}$$

Hence, C is small with $\beta = \lambda(C)\inf_{\boldsymbol{y},\boldsymbol{y}'\in C} f(\boldsymbol{y} - \alpha\boldsymbol{y}')$ and $\mu = \frac{\lambda(\cdot\cap C)}{\lambda(C)}$. Similarly, for any $\boldsymbol{x} \in \mathbb{R}^D$,

$$P(\boldsymbol{x}, C) = \int_C f(\boldsymbol{y} - \alpha\boldsymbol{x})\mathrm{d}\boldsymbol{y} \geq \inf_{\boldsymbol{y}\in C} f(\boldsymbol{y} - \alpha\boldsymbol{x}) \times \lambda(C) > 0.$$

Hence the AR(1) process satisfies Assumption 7.1.2.

Exercise 7.1.1 (Irreducibility of random-walk MH)**.** *Show that the random-walk MH algorithm (Example 6.3.3) on $\mathcal{A} = \mathbb{R}^D$ satisfies Assumption 7.1.2 if q_0 is continuous and everywhere positive and h_π is continuous.*

[1] This definition is a little informal. More precisely, λ is the Lebesgue measure.

Theorem 7.1.4 (Markov chain law of large numbers)**.** *If Assumption 7.1.2 holds and P has an invariant distribution π, then there is a set $S \subseteq \mathcal{A}$ satisfying $\pi(S) = 1$ such that for all $\boldsymbol{x} \in S$ and all $\phi : \mathcal{A} \to \mathbb{R}$ such that $\pi(\phi) < \infty$, the Markov chain $\boldsymbol{X}$ with initial distribution $\delta_{\boldsymbol{x}}$ and transition kernel P satisfies*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{\ell=0}^{k-1} \phi(\boldsymbol{X}_\ell) = \pi(\phi) \quad a.s.$$

Remark 7.1.5 (Ergodicity)**.** *With a little additional work, one can also show that, under the conditions of Theorem 7.1.4, for $\boldsymbol{x} \in S$ and $A \in \mathcal{P}(\mathcal{A})$, $\lim_{k\to\infty} P^k(\boldsymbol{x}, A) = \pi(A)$. A Markov chain satisfying this condition is said to be* **ergodic***.*

7.2 Proof of Markov Chain Law of Large Numbers*

The proof of Theorem 7.1.4 is rather involved, so we start with an overview. The main idea is to construct a Markov chain $\{(\boldsymbol{X}_k, Y_k)\}_{k\in\mathbb{N}}$ on the extended state space $\mathcal{A} \times \{0, 1\}$ such that blocks of the chain marked by $Y_k = 1$ are independent and identically distributed. Define the *first regeneration time* $T_0 := \min\{k \in \mathbb{N} : Y_k = 1\}$ and the *successive regeneration times* $T_i := \min\{k > T_{i-1} : Y_k = 1\}$ $(i = 1, 2, 3, \dots)$. Hence the random variables $\zeta_i(\phi) := \sum_{k=T_{i-1}}^{T_i-1} \phi(\boldsymbol{X}_k)$ are i.i.d.. This i.i.d. structure will (eventually) enable us to apply the standard strong LLN. However, a few facts must be checked first:

1. We must confirm that $\mathbb{P}(T_i < \infty) = 1$ for all $i \in \mathbb{N}$, and in particular that $\mathbb{P}(T_0 < \infty) = 1$ and that $\mathbb{P}(T_i - T_{i-1} < \infty) = 1$. If not, we may never realize the infinite sequence of random variables $\{\zeta_i(\phi)\}_{i\in\mathbb{N}}$.
2. We must show that $\mathbb{E}\{\zeta_i(\phi)\} = Z\pi(\phi)$, where $Z = \mathbb{E}(T_i - T_{i-1})$.
3. Finally, we must verify that the invariant distribution is unique.

We now describe each step in further detail, although some parts we will leave as sketches.

Step 1: extended Markov chain construction. Let $s(\boldsymbol{x}) := \beta \mathbb{1}(\boldsymbol{x} \in C)$ so $P(\boldsymbol{x}, A) \geq s(\boldsymbol{x})\mu(A)$ for all $\boldsymbol{x} \in \mathcal{A}$ and $A \in \mathcal{P}(\mathcal{A})$. Hence, we can define the (homogenous) transition probabilities for the extended Markov chain to be

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{X}_k \in A, Y_k = 1 \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_{k-1}) = s(\boldsymbol{x})\mu(A) \\
&\mathbb{P}(\boldsymbol{X}_k \in A, Y_k = 0 \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_{k-1}) = \tilde{P}(\boldsymbol{x}, A) := P(\boldsymbol{x}, A) - s(\boldsymbol{x})\mu(A).
\end{aligned}$$

Note that by construction $\tilde{P}(\boldsymbol{x}, A) \geq 0$, so the transition probabilities are valid. Moreover, the marginal law of the $\boldsymbol{X}_k$ component of the Markov chain is the same as the original Markov chain since

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{X}_k \in A \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}) \\
&= \mathbb{P}(\boldsymbol{X}_k \in A, Y_k = 1 \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_{k-1}) + \mathbb{P}(\boldsymbol{X}_k \in A, Y_k = 0 \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_{k-1}) \\
&= s(\boldsymbol{x})\mu(A) + \tilde{P}(\boldsymbol{x}, A) \\
&= P(\boldsymbol{x}, A).
\end{aligned}$$

On the other hand, the transition probabilities of the Y_k marginally satisfy

$$\begin{aligned}
&\mathbb{P}(Y_k = 1 \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_{k-1}) = s(\boldsymbol{x}) \\
&\mathbb{P}(Y_k = 0 \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_{k-1}) = 1 - s(\boldsymbol{x}).
\end{aligned} \tag{7.1}$$

Note that $s(\boldsymbol{x}) = 0$ unless $\boldsymbol{x} \in C$, in which case $s(\boldsymbol{x}) = \beta$. Thus, when $\boldsymbol{X}_{k-1}$ is in the small set, with probability β the chain "resets," with the new state $\boldsymbol{X}_k$ distributed according to μ *independent of* $\boldsymbol{X}_{k-1}$. This resampling event is recorded by setting $Y_k = 1$, so

$$\mathbb{P}(\boldsymbol{X}_k \in A \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_k = 1) = \mu(A). \tag{7.2}$$

Finally, we will always take $Y_0 = 0$. We will write $\mathbb{P}_\eta$ to specify that $\mathcal{L}_(\boldsymbol{X}_0) = \eta$ when necessary.

Step 2: characterizing the regeneration times. First, observe that the distribution of T_0 under $\mathbb{P}_\mu$ is the same as $T_i - T_{i-1}$ under $\mathbb{P}$ [for any choice of $\mathcal{L}_{\boldsymbol{X}_0}$]. Now,

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{X}_k \in A, T_0 > k \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, T_0 > k-1) \\
&= \mathbb{P}(\boldsymbol{X}_k \in A, Y_k = 0 \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_{k-1} = 0, \ldots, Y_0 = 0) \\
&= \mathbb{P}(\boldsymbol{X}_k \in A, Y_k = 0 \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_{k-1} = 0) \\
&= \tilde{P}(\boldsymbol{x}, A),
\end{aligned}$$

so we can conclude that

$$\mathbb{P}_\mu(\boldsymbol{X}_k \in A, T_0 > k) = \mu\tilde{P}^k(A). \tag{7.3}$$

Define the measure ν by

$$\nu(A) = \sum_{k=0}^{\infty} \mu\tilde{P}^k(A),$$

where $\mu\tilde{P}^0 := \mu$. Using Eq. (7.3), we can now conclude that

$$\begin{aligned} Z := \mathbb{E}(T_i - T_{i-1}) = \mathbb{E}_\mu(T_0) &= \sum_{k=0}^{\infty} \mathbb{P}_\mu(T_0 > k) \\ &= \sum_{k=0}^{\infty} \mathbb{P}_\mu(\boldsymbol{X}_k \in \mathcal{A}, T_0 > k) \\ &= \sum_{k=0}^{\infty} \mu\tilde{P}^k(\mathcal{A}) \\ &= \nu(\mathcal{A}). \end{aligned}$$

Using similar reasoning, we also have that

$$\mathbb{E}\{\zeta_i(\phi)\} = \mathbb{E}\left\{\sum_{k=T_{i-1}}^{T_i-1} \phi(\boldsymbol{X}_k)\right\} = \mathbb{E}_\mu\left\{\sum_{k=0}^{T_0} \phi(\boldsymbol{X}_k)\right\} = \int \phi(\boldsymbol{x})\nu(\mathrm{d}\boldsymbol{x}). \tag{7.4}$$

Finally, using Eqs. (7.1) and (7.3) we have that

$$\begin{aligned} \mathbb{P}_\mu(T_0 = k) &= \int \mathbb{P}(Y_k = 1 \mid \boldsymbol{X}_{k-1} = \boldsymbol{x}, Y_{k-1} = 0)\mu\tilde{P}^{k-1}(\mathrm{d}\boldsymbol{x}) \\ &= \int s(\boldsymbol{x})\mu\tilde{P}^{k-1}(\mathrm{d}\boldsymbol{x}) \\ &= \mu\tilde{P}^{k-1}(s), \end{aligned}$$

so

$$\mathbb{P}_\mu(T_0 < \infty) = \sum_{k=1}^{\infty} \mu\tilde{P}^{k-1}(s) = \nu(s).$$

Step 3: finiteness of $T_i - T_{i-1}$. Let $L_k := \max\{\ell \in [k] : Y_\ell = 1\}$, so the event $\{T_0 \le k\} = \bigcup_{\ell=1}^{k}\{L_k = \ell\}$. Also, the marginal distribution of Y_k is

$$\mathbb{P}_\eta(Y_k = 1) = \eta P^{k-1}(s). \tag{7.5}$$

We have the identity

$$\begin{aligned}
\mathbb{P}_\pi(\boldsymbol{X}_k \in A, T_0 \le k) &= \sum_{\ell=1}^{k} \mathbb{P}_\pi(\boldsymbol{X}_k \in A, L_k = \ell) \\
&\overset{(i)}{=} \sum_{\ell=1}^{k} \mathbb{P}_\pi(\boldsymbol{X}_k \in A, Y_\ell = 1, Y_{\ell+1} = 0, \ldots, Y_k = 0) \\
&\overset{(ii)}{=} \sum_{\ell=1}^{k} \mathbb{P}_\pi(Y_\ell = 1) \times \mathbb{P}(\boldsymbol{X}_k \in A, Y_{\ell+1} = 0, \ldots, Y_k = 0 \mid Y_\ell = 1) \\
&\overset{(iii)}{=} \sum_{\ell=1}^{k} \pi P^\ell(s) \times \mathbb{P}(\boldsymbol{X}_{k-\ell} \in A, Y_1 = 0, \ldots, Y_{k-\ell} = 0 \mid Y_0 = 1) \\
&\overset{(iv)}{=} \sum_{\ell=1}^{k} \pi P^\ell(s) \times \mathbb{P}_\mu(\boldsymbol{X}_{k-\ell} \in A, T_0 > k - \ell) \\
&\overset{(v)}{=} \pi(s) \sum_{\ell=1}^{k} \mu \tilde{P}^{k-\ell}(A),
\end{aligned}$$

where (i) follows from the definition of L_k, (ii) follows from the definition of conditional probability, (iii) follows from homogeneity of the Markov chain and Eq. (7.5), (iv) follows from Eq. (7.2) and the definition of T_0, and (iv) follows from π being an invariant distribution for a Markov chain with transition kernel P and Eq. (7.3). Hence, after replacing ℓ with $k - \ell$, we have

$$\mathbb{P}_\pi(\boldsymbol{X}_k \in A) = \mathbb{P}_\pi(\boldsymbol{X}_k \in A, T_0 > k) + \pi(s) \sum_{\ell=0}^{k-1} \mu \tilde{P}^\ell(A).$$

Letting $A = \mathcal{A}$ and then taking the limit $k \to \infty$ and using the definition of ν, we obtain

$$\begin{aligned}
1 &= \mathbb{P}_\pi(T_0 > k) + \pi(s) \sum_{\ell=0}^{k-1} \mu \tilde{P}^\ell(\mathcal{A}) \\
1 &= \mathbb{P}_\pi(T_0 = \infty) + \pi(s)\nu(\mathcal{A}). \qquad (7.6)
\end{aligned}$$

Thus, $\nu(\mathcal{A}) = Z = \mathbb{E}_\mu[T_0]$ must be finite since $\beta > 0$ and, by Assumption 7.1.2,

$$\pi(s) = \beta\pi(C) = \beta \sum_{k}^{\infty} 2^{-k} \pi P^k(C) > 0. \qquad (7.7)$$

But if $\mathbb{E}_\mu(T_0) < \infty$, then $\mathbb{P}_\mu(T_0 < \infty) = 1$, and so in addition $\nu(s) = \mathbb{P}_\mu(T_0 < \infty) = \mathbb{P}(T_i - T_{i-1} < \infty) = 1$.

Step 4: invariance of ν. Using the definitions of ν and $\tilde{P}$, and the fact that $\nu(s) = 1$, one can show that $\nu = \nu P$, which means that ν/Z is an invariant distribution of P. An argument by contradiction using Eq. (7.7) shows that, in fact, $\nu/Z = \pi$.

Step 5: finiteness of T_0. Now using Eq. (7.6) and the fact that $\pi(s)\nu(\mathcal{A}) = \nu(s) = 1$, we can conclude that

$$0 = \mathbb{P}_\pi(T_0 = \infty) = \int \mathbb{P}[T_0 = \infty \mid \boldsymbol{X}_0 = \boldsymbol{x}]\pi(\mathrm{d}\boldsymbol{x}).$$

Therefore, for some set S with $\pi(S) = 1$, we must have $\mathbb{P}[T_0 = \infty \mid \boldsymbol{X}_0 = \boldsymbol{x}] = 0$ for $\boldsymbol{x} \in S$.

Step 6: final result. Let $i(k) := \max\{i \in \mathbb{N} : T_i \leq k\}$ denote the number of recurrences after the first up to time k. Putting everything together, it follows from the strong LLN for i.i.d. sequences that $\tilde{\pi}_k(\phi) := \frac{1}{Zi(k)} \sum_{i=1}^{i(k)} \zeta_i(\phi) \overset{a.s.}{\to} \pi(\phi)$. However, $\tilde{\pi}_k(\phi)$ differs from the quantity of interest, $\hat{\pi}_k(\phi) := \frac{1}{k} \sum_{\ell=0}^{k-1} \phi(\boldsymbol{X}_\ell)$, in two ways. First, the normalization in the latter equal the number of terms while it is equal to the expected number of terms in the former. This can be dealt with in the limit since $Zi(k)/k \overset{a.s.}{\to} 1$. Second, $\hat{\pi}_k(\phi)$ contains the terms $k^{-1} \sum_{\ell=1}^{T_0} \phi(\boldsymbol{X}_\ell)$ and $k^{-1} \sum_{\ell=T_{i(k)}}^{k} \phi(\boldsymbol{X}_\ell)$. However, these two terms are asymptotically negligible because of the finiteness of T_0 and $T_{i(k)} - k < T_{i(k)+1} - T_{i(k)}$, which we have shown is finite with probability 1.

Exercise 7.2.1 (Conditional marginal distributions of the extended Markov chain)**.** *Verify (a) Eq.* (7.1) *and (b) Eq.* (7.2)*.*

Exercise 7.2.2 (Expectation of a single block)**.** *Verify the final equality in Eq.* (7.4)*, that* $\mathbb{E}_\mu\{\sum_{k=0}^{T_0} \phi(\boldsymbol{X}_k)\} = \int \phi(\boldsymbol{x})\nu(\mathrm{d}\boldsymbol{x})$*.*

Exercise 7.2.3 (Marginal distribution of Y_k)**.** *Verify Eq.* (7.5)*.*

Exercise 7.2.4 (ν is an invariant measure)**.** *Verify that $\nu = \nu P$. [Hint: use the fact that*

$$\nu(A) = \mu(A) + \sum_{k=0}^{\infty} (\mu \tilde{P}^k) \tilde{P}(A)$$

and recognize that the right-hand side can be rewritten using ν.]

Chapter 8

Lyapunov Functions

With Applications to the Geometric Ergodicity of Markov Chains

8.1 Geometric Ergodicity

Chapter 7 concerns conditions under which the Markov chain "time average" $\frac{1}{k}\sum_{\ell=0}^{k-1}\phi(\boldsymbol{X}_\ell)$ converges to the desired "space average" $\int \phi(\boldsymbol{x})\pi(\mathrm{d}\boldsymbol{x})$, where π is the invariant distribution of the Markov chain. In this chapter we explore a complementary perspective: characterizing the speed of convergence of the *marginal distribution* of $\boldsymbol{X}_k$ to π. Returning to the theme of Chapter 5, we will require a notion of distance between distributions. In keeping with our desire to estimate expectations, we first define a class of functions of interest whose growth is measured relative to a "scaling" function.

Definition 8.1.1 (V-norm)**.** *Given a function $V : \mathcal{A} \to \mathbb{R}_+$, the V-**norm** of a function $\phi : \mathcal{A} \to \mathbb{R}$ is defined as*

$$\|\phi\|_V := \sup_{\boldsymbol{x}\in\mathcal{A}} \frac{|\phi(\boldsymbol{x})|}{1+V(\boldsymbol{x})}.$$

Then, we define the distance between two distributions as the maximum difference in the expectations among all functions with V-norm bounded by 1.

Definition 8.1.2. *The V-**total variational distance** between probability*

measures ν_1 and ν_2 is given by

$$d_V(\nu_1, \nu_2) = \sup_{\substack{\phi:\mathcal{A}\to\mathbb{R} \\ s.t.\ \|\phi\|_V \leq 1}} |\nu_1(\phi) - \nu_2(\phi)|.$$

We will show exponential convergence is this distance under a quantitative version of Assumption 7.1.2 involving V, which in this context is usually called a ***Lyapunov function***.

Assumption 8.1.3. *There exists a Lyapunov function $V : \mathcal{A} \to \mathbb{R}_+$ satisfying the following properties:*

1. *There exist constants $a \in (0, 1)$ and $K \geq 0$ such that, for all $\boldsymbol{x} \in \mathcal{A}$,*

$$(PV)(\boldsymbol{x}) - V(\boldsymbol{x}) \leq -aV(\boldsymbol{x}) + K.$$

2. *There exists a constant $R > 2K/a$ such that $C := \{\boldsymbol{x} \in \mathcal{A} : V(\boldsymbol{x}) \leq R\}$ is a small set.*

We can interpret the first part of the assumption as requiring the expected value of V to decrease after the Markov chain transitions, at least when $V(\boldsymbol{x})$ satisfies $-aV(\boldsymbol{x}) + K < 0$; that is, when $V(\boldsymbol{x}) > K/a$. This latter condition leads to the second part of the assumption requiring that the small set contain all $\boldsymbol{x} \in \mathcal{A}$ such that $V(\boldsymbol{x})$ is smaller than twice the critical value K/a.

Theorem 8.1.4 (Geometric ergodicity)**.** *If Assumption 8.1.3 holds, then P admits a unique invariant distribution π and there exists a constant $B > 0$ such that for $\overline{K} := \max(1, K)$,*

$$\rho := 1 - \beta \min\left\{1/2, \frac{aR - 2\overline{K}}{\beta R + 4\overline{K}}\right\} \in (0, 1),$$

and any probability measure ν,

$$d_V(\nu P^k, \pi) \leq B\rho^k d_V(\nu, \pi). \tag{8.1}$$

A Markov chain satisfying inequality Eq. (8.1) is said to be ***geometrically ergodic***.

Remark 8.1.5. *How ρ depends on the constants involved in Assumption 8.1.3 makes some intuitive sense. The rate of convergence can be faster if β, which measures the "uniformity" of the transition probabilities in the small set, is large. On the other hand, it will also be faster as (i) the convergence rate constant a increases, (ii) the "relaxation" constant K decreases, and (iii) the size of the small set, as measured by R, increases.*

Remark 8.1.6. *Note that Assumption 8.1.3 implies Assumption 7.1.2. It follows from an induction that*

$$(P^k V)(\boldsymbol{x}) \leq (1-a)^k V(\boldsymbol{x}) + K \sum_{\ell=0}^{k-1} (1-a)^\ell \leq (1-a)^k V(\boldsymbol{x}) + K/a.$$

Hence, as long as $V(\boldsymbol{x}) < \infty$, for $k(\boldsymbol{x})$ sufficiently large, $(P^{k(\boldsymbol{x})}V)(\boldsymbol{x}) < R$. But then $P^{k(\boldsymbol{x})}(\boldsymbol{x}, C) > 0$ since $V(\boldsymbol{x}) > R$ for $\boldsymbol{x} \notin C$. Of course, Theorem 8.1.4 guarantees the existence of the invariant distribution. Thus, if Assumption 8.1.3 holds then Theorem 7.1.4 holds as well.

Remark 8.1.7. *In the setting of Chapter 4, let P_k denote the denote the transition kernel for the kth iteration of SGD and let $V(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_\star\|_2^2$. Using the Markov property, we can write $\mathbb{E}_{k-1}(E_k) = P_k V(\boldsymbol{x}_{k-1})$. Therefore, we can rewrite the conclusion of Lemma 4.5.5 as*

$$P_k V(\boldsymbol{x}) \leq \{1 - 2\eta_k \mu(1 - \eta_k L)\} V(\boldsymbol{x}) + 2\eta_k^2 \sigma^2.$$

In other words, under the hypotheses of Lemma 4.5.5, P_k satisfies Assumption 8.1.3(1) with $a = 2\eta_k\mu(1-\eta_k L)$ and $K = 2\eta_k^2\sigma^2$. However we should not expect it to satisfy Assumption 8.1.3(2): due to the discrete nature of the SGD noise, usually the support of $P_k(\boldsymbol{x}, \cdot)$ and $P_k(\boldsymbol{y}, \cdot)$ will be disjoint for $\boldsymbol{x} \neq \boldsymbol{y}$, hence a small set does not exist.

Example 8.1.8 (Geometric ergodicity of the Gaussian AR(1) process)**.** *Consider the univariate version of the Gaussian AR(1) process from Example 3.3.3 given by*

$$X_k = \alpha X_{k-1} + \varepsilon_k,$$

where $\varepsilon_k \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$. Let f denote the density of $\boldsymbol{\varepsilon}_k$ and choose $V(x) = 1/f(bx)$ for a constant $b > 0$ to be determined later. We have

$$PV(x) = \int f(y - \alpha x)/f(by) \mathrm{d}y = \sqrt{\frac{2\pi\sigma_\varepsilon^2}{1-b^2}}\, e^{\frac{\alpha^2 b^2 x^2}{2\sigma_\varepsilon^2(1-b^2)}}.$$

We want the exponential to grow more slowly then $V(x)$, so set $b^2 = 1-\alpha^2/s > 0$ for some $s < 1$. Hence, $\alpha^2/(1-b^2) = (s-1)+1$, so

$$PV(x) = \frac{e^{-\frac{b^2(1-s)x^2}{2\sigma_\varepsilon^2}}}{\sqrt{1-b^2}} V(x).$$

We must now determine how large x must be to ensure the term multiplying $V(x)$ is less than some fixed $\gamma \in (0,1)$. So, solving

$$\frac{e^{-\frac{b^2(1-s)x^2}{2\sigma_\varepsilon^2}}}{\sqrt{1-b^2}} \leq \gamma$$

for x^2 we obtain

$$x^2 \geq x_\star^2 := \frac{\sigma_\varepsilon^2 \log \frac{1-b^2}{\gamma^2}}{(1-s)b^2}.$$

Thus, if $x^2 \geq x_*^2$, then $PV(x) \leq \gamma V(x)$, while if $x^2 < x_*^2$, then

$$PV(x) \leq \frac{V(x)}{\sqrt{1-b^2}} \leq \gamma V(x) + \left(\frac{1}{\sqrt{1-b^2}} - \gamma\right) V(x_\star).$$

We can now conclude that Assumption 8.1.3(1) holds with $a = 1-\gamma$ and $K = (\frac{1}{\sqrt{1-b^2}} - \gamma)V(x_*)$. Since the set $C = \{x \in \mathbb{R} : V(x) \leq R\}$ is in the form of an interval and f is continuous and everywhere positive, it follows from Example 7.1.3 that Assumption 8.1.3(2) holds as well.

Remark 8.1.9. *Recall from Example 3.3.3 that the invariant distribution of the Gaussian AR(1) process is $\pi = \mathcal{N}(0, \frac{\sigma_\varepsilon^2}{1-\alpha^2})$. So, $V(\boldsymbol{x}) \propto h_\pi(\boldsymbol{x})^{-\delta}$ for some $\delta < 1$. This behavior for the Lyapunov function is quite common. It is also essentially the "best case scenario" since, if $V(\boldsymbol{x}) \propto h_\pi(\boldsymbol{x})^{-1}$, then $\pi(V) = \infty$, which contradicts Theorem 8.1.4.*

8.2 Geometric Ergodicity of the Random-walk Metropolis–Hastings Algorithm

The conditions required for MCMC algorithms to be ergodic in the sense of Theorem 7.1.4 are quite weak. Exercise 7.1.1 provides one illustrative example, where all that is required in the case of the random-walk MH

algorithm (Example 6.3.3) is that the target density is continuous. The conditions for geometric ergodicity are substantially stronger. To get a sense of how much stronger, we continue to focus on the random-walk MH algorithm. The following, very restricted result, is representative.

Proposition 8.2.1. *For $\mathcal{A} = \mathbb{R}_+$ and $h_\pi(x) = \lambda^{-1}e^{-\lambda x}$, the random-walk MH algorithm satisfies Assumption 8.1.3 if q_0 is continuous and everywhere positive.*

Proof. For some $\delta \in (0,1)$, let $V(x) = e^{\delta\lambda x} \propto h_\pi(x)^{-\delta}$ For $x > 0$, we have

$$\begin{aligned}(PV)(x) = &\int_0^x V(y)q_0(x-y)\mathrm{d}y + \int_x^\infty \frac{h_\pi(y)}{h_\pi(x)}V(y)q_0(x-y)\mathrm{d}y \\ &+ V(x)\int_x^\infty \left\{1 - \frac{h_\pi(y)}{h_\pi(x)}\right\} q_0(x-y)\mathrm{d}y,\end{aligned}$$

where the first integral is for the case when the proposal is less than x, in which case it is accepted, the second (third) integral is for the case when the proposal is greater than x and accepted (rejected). Making the change of variable $y \leftarrow 2x - y$ and using the symmetry of q_0, the first integral can be rewritten as $\int_x^{2x} V(2x-y)q_0(x-y)\mathrm{d}y$. Thus,

$$(PV)(x) \le \int_x^\infty I(x,y)q_0(x-y)\mathrm{d}y,$$

where, letting $z = y - x$,

$$\begin{aligned}I(x,y) &= V(2x-y) + \frac{h_\pi(y)}{h_\pi(x)}V(y) + V(x)\left\{1 - \frac{h_\pi(y)}{h_\pi(x)}\right\} \\ &= e^{\delta\lambda(2x-y)} + e^{\lambda(x-y)+\delta\lambda y} + e^{\delta\lambda x}(1 - e^{\lambda(x-y)}) \\ &= e^{\delta\lambda x}\left\{e^{-\delta\lambda z} + e^{(\delta-1)\lambda z} + 1 - e^{-\lambda z}\right\} \\ &= e^{\delta\lambda x}\left\{2 - (1 - e^{(\delta-1)\lambda z})(1 - e^{-\lambda z})\right\}\end{aligned}$$

Hence, for any $z^* > 0$ and $\epsilon := (1 - e^{(\delta-1)\lambda z^*})(1 - e^{-\lambda z^*})$,

$$I(x,y) \le \begin{cases} 2V(x) & \text{for } y \ge x \\ (2-\epsilon)V(x) & \text{for } y \ge x + z^*. \end{cases}$$

Since $\delta < 1$, we know that $\epsilon \in (0, 1)$ and since q_0 is symmetric, positive, and continuous, $p^* := \int_{z^*}^{\infty} q_0(y)\mathrm{d}y < 1/2$. Thus, we conclude that

$$\begin{aligned}(PV)(x) &\leq 2V(x)\int_x^{x+z^*} q_0(x-y)\mathrm{d}y + (2-\epsilon)V(x)\int_{x+z^*}^{\infty} q_0(x-y)\mathrm{d}y \\ &= (1-\epsilon^*)V(x),\end{aligned}$$

where $\epsilon^* = 2\epsilon p^* < 1$. Thus, Assumption 8.1.3(1) is satisfied with $a = \epsilon^*$ and $K = 0$. Moreover, one can check that $[0, R]$ is a small set for any $R > 0$, so Assumption 8.1.3(2) is satisfied as well. □

Focusing on the case of $\mathcal{A} = \mathbb{R}$ for simplicity, the above logic can be generalized to any target distribution satisfying the following condition.

Assumption 8.2.2 (Log-concavity in the tails)**.** *A density f is said to* **log-concave in the tails** *if there exists $\lambda > 0$ and some $x^* > 0$ such that for all $y > x > x^*$,*

$$\log f(x) - \log f(y) \geq \lambda(y - x)$$

and for $y < x < -x^*$,

$$\log f(x) - \log f(y) \geq \lambda(x - y).$$

Essentially, log-concavity in the tails corresponds to the tails of the density smoothly decaying at least as quickly as the exponential $e^{-\lambda|x|}$, hence generalizing the case considered above.

Theorem 8.2.3. *Assume $h_\pi > 0$ is continuous and log-concave in the tails. If h_π is symmetric, then the random-walk MH algorithm with continuous $q_0 > 0$ satisfies Assumption 8.1.3 with $V(x) = e^{\delta\lambda x}$ for any $\delta \in (0, 1)$. If h_π is not symmetric, then the same conclusion holds if for some $b > 0$, $q_0(x) \leq be^{-\lambda|x|}$.*

Interestingly, a near converse also holds.

Theorem 8.2.4. *Assume $h_\pi > 0$ is continuous and $\int |x|q_0(x)\mathrm{d}x < \infty$. If the random-walk MH algorithm is geometrically ergodic, then there exists $\lambda > 0$ such that $\int h_\pi(x)e^{\lambda|x|}\mathrm{d}x < \infty$.*

In other words, geometric ergodicity of the random-walk MH algorithm implies that the tails of the target density must decay at least exponentially fast.

8.3 Proof of General Geometric Ergodicity Condition*

To prove Theorem 8.1.4, we will rely on a clever trick that is frequently useful. The idea is to work with a different, specially tailored distance that is equivalent to the original distance, in the sense that the ratio of the distances is uniformly bounded away from zero and infinity.

Definition 8.3.1. *Distances d and d' on a set $\mathcal{S}$ are* **equivalent** *if there exist finite constants $0 < c_* \le c^* < \infty$ such that for all $s, t \in \mathcal{S}$,*

$$c_* d'(s,t) \le d(s,t) \le c^* d'(s,t).$$

We will work with a family of equivalent distances parameterized by a positive constant $\sigma > 0$, which is used to define a modified V-norm:

$$\|\phi\|_{V,\sigma} := \sup_{\boldsymbol{x} \in \mathcal{A}} \frac{|\phi(\boldsymbol{x})|}{1 + \sigma V(\boldsymbol{x})},$$
$$d_{V,\sigma}(\nu_1, \nu_2) = \sup_{\phi \,:\, \|\phi\|_{V,\sigma} \le 1} |\nu_1(\phi) - \nu_2(\phi)|.$$

Of course if $\sigma = 1$ then $\|\phi\|_{V,\sigma} = \|\phi\|_V$ and $d_{V,\sigma} = d_V$.

Lemma 8.3.2. *The distances $d_{V,\sigma}$ and d_V are equivalent.*

> **Exercise 8.3.1** (V-norm equivalence)**.** *Prove Lemma 8.3.2. [Hint: consider the cases of $\sigma < 1$ and $\sigma > 1$ separately.]*

We will prove the following result, which will imply Theorem 8.1.4.

Theorem 8.3.3. *If Assumption 8.1.3 holds, then there exists $\rho \in (0,1)$ and $\sigma > 0$ such that for any probability measures ν_1 and ν_2,*

$$d_{V,\sigma}(\nu_1 P, \nu_2 P) \le \rho \, d_{V,\sigma}(\nu_1, \nu_2). \tag{8.2}$$

Moreover, Eq. (8.2) implies that P admits a unique invariant distribution π satisfying $\pi(V) < \infty$.

Exercise 8.3.2 (Exponential convergence of modified distance implies geometric ergodicity)**.** *Show that Theorem 8.3.3 implies Theorem 8.1.4.*

Before proving Theorem 8.3.3, we construct *another* distance that is almost equal to $d_{V,\sigma}(\nu_1, \nu_2)$:

$$\|\phi\|_{V,\sigma,L} := \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in\mathcal{A}\\ \boldsymbol{x}\neq\boldsymbol{y}}} \frac{|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})|}{2 + \sigma V(\boldsymbol{x}) + \sigma V(\boldsymbol{y})},$$
$$\tilde{d}_{V,\sigma}(\nu_1, \nu_2) = \sup_{\phi\,:\,\|\phi\|_{V,\sigma,L}\leq 1} |\nu_1(\phi) - \nu_2(\phi)|.$$

Lemma 8.3.4. *The identity* $\|\phi\|_{V,\sigma,L} = \inf_{c\in\mathbb{R}} \|\phi + c\|_{V,\sigma}$ *holds, so in particular* $d_{V,\sigma}(\nu_1, \nu_2) = \tilde{d}_{V,\sigma}(\nu_1, \nu_2)$.

Proof. To prove the identity, we first observe that for $r(\boldsymbol{x}) := \frac{|\phi(\boldsymbol{y})|}{1+\sigma V(\boldsymbol{y})}$ and $w(\boldsymbol{x}) := \frac{1}{1+\sigma V(\boldsymbol{y})}$, we have

$$\begin{aligned}
\|\phi\|_{V,\sigma,L} &= \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in\mathcal{A}\\ \boldsymbol{x}\neq\boldsymbol{y}}} \frac{|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})|}{2 + \sigma V(\boldsymbol{x}) + \sigma V(\boldsymbol{y})} \\
&\leq \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in\mathcal{A}\\ \boldsymbol{x}\neq\boldsymbol{y}}} \frac{|\phi(\boldsymbol{x})| + |\phi(\boldsymbol{y})|}{2 + \sigma V(\boldsymbol{x}) + \sigma V(\boldsymbol{y})} \\
&= \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{A}} \frac{w(\boldsymbol{y})r(\boldsymbol{x}) + w(\boldsymbol{x})r(\boldsymbol{y})}{w(\boldsymbol{x}) + w(\boldsymbol{y})} \\
&\leq \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{A},\alpha\in[0,1]} \alpha r(\boldsymbol{x}) + (1-\alpha) r(\boldsymbol{y}) \\
&= \sup_{\boldsymbol{x}\in\mathcal{A}} r(\boldsymbol{x}) = \|\phi\|_V.
\end{aligned}$$

Therefore, we have $\|\phi\|_{V,\sigma,L} = \|\phi + c\|_{V,\sigma,L} \leq \|\phi + c\|_V$ and hence $\|\phi\|_{V,\sigma,L} \leq \inf_{c\in\mathbb{R}} \|\phi + c\|_V$.

Now we show the reverse inequality. Since for $a > 0$, $\|a\phi\|_{V,\sigma} = a\|\phi\|_{V,\sigma}$ and $\|a\phi\|_{V,\sigma,L} = a\|\phi\|_{V,\sigma,L}$, without loss of generality assume that $\|\phi\|_{V,\sigma} = 1$. Hence, $\phi(\boldsymbol{x}) \leq |\phi(\boldsymbol{y})| + |\phi(\boldsymbol{x}) - |\phi(\boldsymbol{y})| \leq |\phi(\boldsymbol{y})| + 2 + \sigma V(\boldsymbol{x}) + \sigma V(\boldsymbol{y})$, which we can rewrite at

$$1 + \sigma V(\boldsymbol{x}) - \phi(\boldsymbol{x}) \leq -1 - \sigma V(\boldsymbol{y}) - |\phi(\boldsymbol{y})|.$$

From this inequality and the fact that $V(\boldsymbol{y})$ and $|\phi(\boldsymbol{x})|$ are nonnegative and finite, it follows that $c^* := \inf_{\boldsymbol{x}\in\mathcal{A}} 1 + \sigma V(\boldsymbol{x}) - |\phi(\boldsymbol{x})|$ is bounded from below and thus $|c^*| < \infty$. Next, note that

$$\phi(\boldsymbol{x}) + c^* \le \phi(\boldsymbol{x}) + 1 + \sigma V(\boldsymbol{x}) - \phi(\boldsymbol{x}) = 1 + \sigma V(\boldsymbol{x})$$

and, since $\|\phi\|_{V,\sigma,L} = 1$,

$$\begin{aligned}\phi(\boldsymbol{x}) + c^* &\ge \inf_{\boldsymbol{y}\in\mathcal{A}} \phi(\boldsymbol{x}) + 1 + \sigma V(\boldsymbol{y}) - \phi(\boldsymbol{y}) \\ &\ge \inf_{\boldsymbol{y}\in\mathcal{A}} 1 + \sigma V(\boldsymbol{y}) - \|\phi\|_{V,\sigma,L}\{2 + \sigma V(\boldsymbol{x}) + \sigma V(\boldsymbol{y})\} \\ &\ge -\{1 + \sigma V(\boldsymbol{x})\}.\end{aligned}$$

Hence $|\phi(\boldsymbol{x}) + c^*| \le 1 + \sigma V(\boldsymbol{x})$ and

$$\inf_{c\in\mathbb{R}} \|\phi + c\|_{V,\sigma} = \inf_{c\in\mathbb{R}} \sup_{\boldsymbol{x}\in\mathcal{A}} \frac{|\phi(\boldsymbol{x}) + c|}{1 + \sigma V(\boldsymbol{x})} \le \sup_{\boldsymbol{x}\in\mathcal{A}} \frac{|\phi(\boldsymbol{x}) + c^*|}{1 + \sigma V(\boldsymbol{x})} \le 1 = \|\phi\|_{V,\sigma,L},$$

confirming that $\|\phi\|_{V,\sigma,L} = \inf_{c\in\mathbb{R}} \|\phi + c\|_V$. This identity implies $\{\phi : \|\phi\|_{V,\sigma,L} \le 1\} = \{\phi + c : \|\phi\|_{V,\sigma} \le 1, c \in \mathbb{R}\}$, so $d_{V,\sigma}(\nu_1, \nu_2) = \tilde{d}_{V,\sigma}(\nu_1, \nu_2)$. □

Proof of Theorem 8.3.3. If $\|P\phi\|_{V,\sigma,L} \le \rho\|\phi\|_{V,\sigma,L}$, the result will follow from Lemma 8.3.4 and the bound

$$\begin{aligned}d_{V,\sigma}(\nu_1 P, \nu_2 P) &= \sup_{\phi\colon \|\phi\|_{V,\sigma}\le 1} \int \phi(\boldsymbol{x})(\nu_1 P)(\mathrm{d}\boldsymbol{x}) - \int \phi(\boldsymbol{x})(\nu_2 P)(\mathrm{d}\boldsymbol{x}) \\ &= \sup_{\phi\colon \|\phi\|_{V,\sigma}\le 1} \int\int \phi(\boldsymbol{x}) P(\boldsymbol{y}, \mathrm{d}\boldsymbol{x})\nu_1(\mathrm{d}\boldsymbol{y}) - \int\int \phi(\boldsymbol{x}) P(\boldsymbol{y}, \mathrm{d}\boldsymbol{x})\nu_2(\mathrm{d}\boldsymbol{y}) \\ &= \sup_{\phi\colon \|\phi\|_{V,\sigma}\le 1} \int (P\phi)(\boldsymbol{x})\nu_1(\mathrm{d}\boldsymbol{x}) - \int (P\phi)(\boldsymbol{x})\nu_2(\mathrm{d}\boldsymbol{x}) \\ &\le \rho \sup_{\phi\colon \|\phi\|_{V,\sigma}\le 1} \int \phi(\boldsymbol{x})\nu_1(\mathrm{d}\boldsymbol{x}) - \int \phi(\boldsymbol{x})\nu_2(\mathrm{d}\boldsymbol{x}) \\ &= \rho\, d_{V,\sigma}(\nu_1, \nu_2).\end{aligned}$$

We will fix a test function ϕ that, without loss of generality, satisfies $\|\phi\|_{V,\sigma,L} = 1$. By Lemma 8.3.4, without loss of generality we can assume $\|\phi\|_{V,\sigma} = 1$ (by possibly replacing ϕ by $\phi + c$). Moreover, in this case we can rewrite the conclusion of the theorem as requiring that, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}$,

$$|P\phi(\boldsymbol{x}) - P\phi(\boldsymbol{y})| \le \rho\{2 + \sigma V(\boldsymbol{x}) + \sigma V(\boldsymbol{y})\}. \tag{8.3}$$

Therefore, we turn to proving Eq. (8.3). Clearly the inequality holds if $\boldsymbol{x} = \boldsymbol{y}$, so assume $\boldsymbol{x} \neq \boldsymbol{y}$. We divide the verification of Eq. (8.3) into two cases depending on the value of $V(\boldsymbol{x}) + V(\boldsymbol{y})$.

Case 1: $V(\boldsymbol{x})+V(\boldsymbol{y}) \geq R$**.** In this case we essentially rely on the contraction property from Assumption 8.1.3(1). We have the bound

$$\begin{aligned} |P\phi(\boldsymbol{x}) - P\phi(\boldsymbol{y})| &\leq P|\phi|(\boldsymbol{x}) + P|\phi|(\boldsymbol{y}) \\ &\overset{(i)}{\leq} P(1+\sigma V)(\boldsymbol{x}) + P(1+\sigma V)(\boldsymbol{y}) \\ &= 2 + \sigma PV(\boldsymbol{x}) + \sigma PV(\boldsymbol{y}) \\ &\overset{(ii)}{\leq} 2 + \sigma\gamma V(\boldsymbol{x}) + \sigma\gamma V(\boldsymbol{y}) + 2\sigma K, \end{aligned}$$

where (i) follows from having $\|\phi\|_{V,\sigma,L} \leq 1$ and (ii) follows from Assumption 8.1.3(1). Let $\gamma_0 \in (\gamma, 1)$ be a constant to be determined shortly and rewrite the final bound as

$$\begin{aligned} 2 + \sigma\gamma_0 V(\boldsymbol{x}) + \sigma\gamma_0 V(\boldsymbol{y}) \; - \; \sigma(\gamma_0 - \gamma)V(\boldsymbol{x}) - \sigma(\gamma_0 - \gamma)V(\boldsymbol{y}) + 2\sigma K \\ \leq 2 + \sigma\gamma_0 V(\boldsymbol{x}) + \sigma\gamma_0 V(\boldsymbol{y}) - \sigma(\gamma_0 - \gamma)R + 2\sigma K, \end{aligned}$$

where we have used the fact that $V(\boldsymbol{x}) + V(\boldsymbol{y}) \geq R$. Now, to obtain the desired upper bound, we first choose γ_0 such that $-\sigma(\gamma_0 - \gamma)R + 2\sigma K = 0$, which leads to $\gamma_0 = \gamma + 2K/R$. Note that $\gamma_0 < 1$ since $R > 2K/(1-\gamma)$. For some $\gamma_1 \in (\gamma_0, 1)$, we now rewrite the upper bound again as

$$\begin{aligned} &2 + \sigma\gamma_0 V(\boldsymbol{x}) + \sigma\gamma_0 V(\boldsymbol{y}) \\ &= \gamma_1\{2 + \sigma V(\boldsymbol{x}) + \sigma V(\boldsymbol{y})\} + \underbrace{2(1-\gamma_1) - \sigma(\gamma_1 - \gamma_0)\{V(\boldsymbol{x}) + V(\boldsymbol{y})\}}_{(\star)} \end{aligned}$$

and choose γ_1 such that $(\star) \leq 0$. Since $V(\boldsymbol{x}) + V(\boldsymbol{y}) \geq R$, it suffices to choose γ_1 such that $2(1-\gamma_1) - \sigma(\gamma_1 - \gamma_0)R = 0$, which yields $\gamma_1 := (2 + \sigma R\gamma_0)/(2 + \sigma R)$. We are guaranteed that $\gamma_1 < 1$ since $\gamma_0 < 1$ and, moreover, $\gamma_1 > \gamma_0$ as long as $\sigma < 1$. Hence, assuming $\sigma < 1$, we conclude that

$$|P\phi(\boldsymbol{x}) - P\phi(\boldsymbol{y})| \leq \gamma_1\{2 + \sigma V(\boldsymbol{x}) + \sigma V(\boldsymbol{y})\}.$$

Case 2: $V(\boldsymbol{x}) + V(\boldsymbol{y}) < R$**.** In this case we must have $\boldsymbol{x}, \boldsymbol{y} \in C$, so we will rely on the small set assumption to ensure contractivity of the transition kernel, as $P(\boldsymbol{x}, \cdot)$ and $P(\boldsymbol{y}, \cdot)$ must be somewhat the same. In particular, by

Assumption 8.1.3, for all $\boldsymbol{x}' \in C$ and $A \in \mathcal{F}$, $\tilde{P}(\boldsymbol{x}', A) := P(\boldsymbol{x}', A) - \beta\mu(A) \geq 0$. Hence,

$$\begin{aligned}
|P\phi(\boldsymbol{x}) - P\phi(\boldsymbol{y})| &= |\tilde{P}\phi(\boldsymbol{x}) - \tilde{P}\phi(\boldsymbol{y})| \\
&\leq \tilde{P}(1+\sigma V)(\boldsymbol{x}) + \tilde{P}(1+\sigma V)(\boldsymbol{y}) \\
&\overset{(i)}{=} 2(1-\beta) + \sigma\tilde{P}V(\boldsymbol{x}) + \sigma\tilde{P}V(\boldsymbol{y}) \\
&\overset{(ii)}{\leq} 2(1-\beta) + \sigma PV(\boldsymbol{x}) + \sigma PV(\boldsymbol{y}) \\
&\overset{(iii)}{\leq} 2(1-\beta) + \gamma\sigma V(\boldsymbol{x}) + \gamma\sigma V(\boldsymbol{y}) + 2\sigma K,
\end{aligned}$$

where first the two steps follow as in Case 1, (i) follows from $\tilde{P}(\boldsymbol{x}, \mathcal{A}) = 1-\beta$, (ii) follows from $\tilde{P}V(\boldsymbol{x}) \leq PV(\boldsymbol{x})$ since V is nonnegative, and (iii) follows from Assumption 8.1.3(1). Now we exploit our freedom to choose σ, which we set to $\sigma = \frac{\beta}{2\max(1,K)} < 1$, so

$$\begin{aligned}
|P\phi(\boldsymbol{x}) - P\phi(\boldsymbol{y})| &\leq 2(1-\beta/2) + \gamma\sigma V(\boldsymbol{x}) + \gamma\sigma V(\boldsymbol{y}) \\
&\leq \gamma_2\{2 + \sigma V(\boldsymbol{x}) + \sigma V(\boldsymbol{y})\},
\end{aligned}$$

where $\gamma_2 = \max(1-\beta/2, \gamma) < 1$. Since $\gamma_1 > \gamma_0 > \gamma$, Eq. (8.2) holds for $\rho = \max(1-\beta/2, \gamma_1) < 1$, which after simplification takes the form given in Theorem 8.1.4.

The conclusion that the Markov chain admits a unique invariant distribution follows from a simple but slightly technical argument. Essentially the results follows by noting that the sequence of probability measures given by $\nu_k := \delta_{\boldsymbol{x}} P^k$ is a ***Cauchy sequence*** under the metric $d_{V,\sigma}$ since $d_{V,\sigma}(\nu_k, \nu_{k+1}) \leq \rho^k d_{V,\sigma}(\nu_1, \delta_{\boldsymbol{x}}) \to 0$ as $k \to \infty$. Since $d_{V,\sigma}$ is complete for the space of probability measures for which integrating V is finite, this implies that there exists a distribution π such that $d_{V,\sigma}(\nu_k, \pi) \to 0$ and $\pi(V) < \infty$. Thus, $P\pi = \lim_{k\to\infty} P\nu_k = \lim_{k\to\infty} \nu_{k+1} = \pi$, so π is an invariant distribution. On the other hand, if there exists a second invariant distribution π', then $d_{V,\sigma}(\pi, \pi') = d_{V,\sigma}(\pi P, \pi' P) \leq \rho d_{V,\sigma}(\pi, \pi')$ so in fact $d_{V,\sigma}(\pi, \pi') = 0$. Thus, the invariant distribution is unique. □

Exercise 8.3.3 (Value of B)**.** *Use your derivations from Exercises 8.3.1 and 8.3.2 to explicitly calculate the value of B.*

Appendix A

Mathematical Background

A.1 L^p Spaces and Inequalities

The final mathematical background we need relates to function spaces and inequalities.

Definition A.1.1. *Given a measure μ and number $p > 0$, define the collection $L^p = L^p(\mu)$ of functions $\phi : \mathcal{A} \to \mathbb{R}$ having finite L^p norm:*

$$\|\phi\|_{L^p} := \{\mu(|\phi|^p)\}^{1/p} < \infty.$$

The case of $p = 2$ is noteworthy as it generalizes the Euclidean norm. Letting $\langle \phi, \psi \rangle := \mu(\phi\psi)$ denote the inner product, we have $\|\phi\|_{L^2}^2 = \langle \phi, \phi \rangle$. The following two results establish bounds on the L^p norms of products and sums of functions.

Lemma A.1.2 (Hölder and Cauchy–Schwarz inequalities)**.** *For all $p, q, r > 0$ satisfying $1/p + 1/q = 1/r$,*

$$\|\phi\psi\|_{L^r} \leq \|\phi\|_{L^p} \|\psi\|_{L^q}.$$

In particular, taking $p = q = 2$ and $r = 1$,

$$|\langle \phi, \psi \rangle| \leq \|\phi\psi\|_{L^1} \leq \|\phi\|_{L^2} \|\psi\|_{L^2}.$$

Lemma A.1.3 (Minkowski inequality)**.** *For all $p \geq 1$,*

$$\|\phi + \psi\|_{L^p}^p \leq \|\phi\|_{L^p}^p + \|\psi\|_{L^p}^p.$$

Thus, L^p norms satisfy a kind of generalized triangle inequality.

Often we wish to ensure a random variable is unlikely to be too large. The major tool for doing so is the following:

Lemma A.1.4 (Markov inequality)**.** *If X is a nonnegative random variable, then for all $t > 0$,*

$$\mathbb{P}\{X > t\mathbb{E}(X)\} \leq \frac{1}{t}.$$

Proof. Without loss of generality assume $\mathbb{E}(X) = 1$, so

$$\begin{aligned} t\mathbb{1}(X > t) &\leq X \\ \mathbb{E}\{t\mathbb{1}(X > t)\} &\leq \mathbb{E}(X) \\ t\mathbb{P}(X > t) &\leq 1. \end{aligned}$$

□

We can always apply the Markov inequality to nonnegative $\phi(X)$ for arbitrary X. For example, if $\phi(X) = (X - \mathbb{E}\{X\})^2$, then we obtain the ***Chebyshev inequality***

$$\mathbb{P}\{|X - \mathbb{E}X| > \varepsilon\} \leq \frac{\operatorname{Var}(X)}{\varepsilon^2}$$

using the substitution $\varepsilon^2 = t \operatorname{Var}(X)$:

$$\mathbb{P}\{(X - \mathbb{E}X)^2 > t \operatorname{Var}(X)\} \leq \frac{1}{t} \quad \Longrightarrow \quad \mathbb{P}\{(X - \mathbb{E}X)^2 > \varepsilon^2\} \leq \frac{\operatorname{Var}(X)}{\varepsilon^2}.$$

A.2 Linear Algebra and Vector Calculus

For column vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D$, let $\langle \boldsymbol{u}, \boldsymbol{v} \rangle := \boldsymbol{u}^\top \boldsymbol{v} = \sum_{i=1}^D u_i v_i$ denote the ***Euclidean inner product*** and $\|\boldsymbol{u}\|_2 := \langle \boldsymbol{u}, \boldsymbol{u} \rangle^{1/2} = (\sum_{i=1}^D u_i^2)^{1/2}$ denote the ***Euclidean (ℓ_2) norm***. A matrix $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ is ***symmetric*** if $\boldsymbol{A} = \boldsymbol{A}^\top$. A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ is ***positive definite (pd)*** if

$$\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u} > 0 \quad \text{for all } \boldsymbol{u} \in \mathbb{R}^D \setminus \{\boldsymbol{0}\}, \tag{A.1}$$

which we denote by $\boldsymbol{A} \succ 0$. It is ***positive semidefinite (psd)*** if Eq. (A.1) holds with a $\geq$ rather than a $>$, which we denote by $\boldsymbol{A} \succeq 0$. More generally, we write $\boldsymbol{A} \succ \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B} \succ 0$ and similarly for $\succeq$.

Any psd matrix $\boldsymbol{A}$ induces an inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{A}} := \boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{v}$ and associated norm $\|\boldsymbol{u}\|_{\boldsymbol{A}} := \langle \boldsymbol{u}, \boldsymbol{u} \rangle_{\boldsymbol{A}}^{1/2}$.

Lemma A.2.1 (Cauchy–Schwarz inequality for vectors)**.** *For any psd matrix matrix $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ and vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D$,*

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{A}} \leq \|\boldsymbol{u}\|_{\boldsymbol{A}} \|\boldsymbol{v}\|_{\boldsymbol{A}}.$$

Define the ***spectral norm*** of a matrix $\boldsymbol{A} \in \mathbb{R}^{D' \times D}$ by

$$\|\boldsymbol{A}\|_2 := \sup_{\boldsymbol{x} \in \mathbb{R}^D \setminus \{\boldsymbol{0}\}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}.$$

So, by construction $\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \|\boldsymbol{A}\|_2 \|\boldsymbol{x}\|_2$. Moreover, for matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, if $\boldsymbol{A}\boldsymbol{B}$ is well-defined, then $\|\boldsymbol{A}\boldsymbol{B}\|_2 \leq \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_2$.

Definition A.2.2 (Orthonormal matrix)**.** *A matrix $\boldsymbol{Q} \in \mathbb{R}^{D \times D}$ is* **orthonormal** *if $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^\top = I$, so in particular $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\top$.*

Theorem A.2.3 (Spectral theorem)**.** *Any symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ can be written as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{Q}$ is orthonormal and $\boldsymbol{\Lambda}$ is diagonal.*

Proposition A.2.4. *A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ is positive definite (positive semidefinite) if and only if $\Lambda_{ii} > 0$ ($\Lambda_{ii} \geq 0$) for all $i = 1, \ldots, D$.*

Proposition A.2.5. *For any matrix $\boldsymbol{A} \in \mathbb{R}^{D' \times D}$, the matrix $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ is positive semidefinite and $\|\boldsymbol{A}\|_2 = \max_i \Lambda_{ii}^{1/2}$*

Definition A.2.6 (Gradient and Hessian)**.** *For a function $\phi : \mathbb{R}^D \to \mathbb{R}$, the* **gradient** *$\nabla \phi : \mathbb{R}^D \to \mathbb{R}^D$ is defined as $\nabla \phi(\boldsymbol{x})_i := \frac{\partial \phi}{\partial \boldsymbol{x}_i}(\boldsymbol{x})$ and the* **Hessian** *$\nabla^2 \phi : \mathbb{R}^D \to \mathbb{R}^{D \times D}$ is defined as $\nabla^2 \phi(\boldsymbol{x})_{ij} := \frac{\partial^2 \phi}{\partial \boldsymbol{x}_i \partial \boldsymbol{x}_j}(\boldsymbol{x})$. We will also use the shorthands $\phi' := \nabla \phi$ and $\phi'' := \nabla^2 \phi$.*

Bibliography

K. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Index