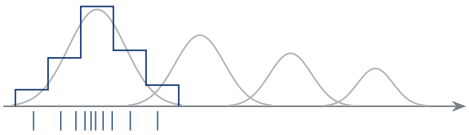
2

Monte Carlo and the Statistics of Returns



The simplest way to learn about a policy is to run it and look at what happens. This chapter develops the statistical tools that turn simulation into estimation with guarantees: importance sampling, which allows one to evaluate a policy from the runs of another; laws of large numbers and central limit theorems for asymptotic confidence intervals; Hoeffding's inequality for finite-sample guarantees; and the Chernoff method, which controls the tails of a random variable through its exponential moments. Exponential moments give rise to the β-means, a family of functionals that will accompany us throughout the book. We then turn to sequential estimation and stochastic gradient descent, to the estimation of a whole distribution rather than of its mean, and finally to censored observations, which arise as soon as decisions hide part of the data.

2.1 Monte Carlo policy evaluation

Consider a finite-horizon MDP, a policy $\pi \in \Pi_{\mathrm{H}}$ and an initial state s. We assume that we can *simulate* the interaction: draw $S_1 = s$, choose actions according to π, and draw transitions according to the kernels—either because they are known or because a generative model in the sense of Proposition 1.12 is available. Running the policy n times independently produces i.i.d. copies $W^{(1)}, \dots, W^{(n)}$ of the return W_H, with common law $\nu^\pi(s)$. The *Monte Carlo estimator* of the value $V^\pi(s) = \mathbb{E}[\nu^\pi(s)]$ is the empirical mean

$$\hat{V}_n(s) \;\coloneqq\; \frac{1}{n}\sum_{i=1}^{n} W^{(i)}.$$

More generally, the empirical measure $\hat{\nu}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{W^{(i)}}$ estimates $\nu^\pi(s)$ itself, and any functional $\psi(\nu^\pi(s))$ is estimated by the *plug-in* estimator $\psi(\hat{\nu}_n)$; the empirical mean is the plug-in estimator of the mean. Algorithm 2.1 summarizes the procedure. Its cost is nH calls to the simulator, and it does not require the state space to be small: the states are never enumerated.

For the rest of the chapter we write $X_1, \dots, X_n$ for i.i.d. real random variables with law ν, mean m and variance σ^2 when they exist, and $\bar{X}_n = \frac{1}{n}\sum_{i\le n} X_i$ for their empirical mean; in the application $X_i = W^{(i)}$, $m = V^\pi(s)$ and $\nu = \nu^\pi(s)$.

Proposition 2.1 (Consistency). If $\mathbb{E}|X_1| < \infty$ then $\bar{X}_n \to m$ almost surely as $n \to \infty$. If moreover $\sigma^2 < \infty$ then $\mathbb{E}[\bar{X}_n] = m$ and $\mathbb{V}\mathrm{ar}[\bar{X}_n] = \sigma^2/n$.

Algorithm 2.1: Monte Carlo policy evaluation

Input: initial state s, policy π, number of runs n, a simulator of the MDP

for $i = 1, \dots, n$ **do**
  $S_1 \leftarrow s$, $w \leftarrow 0$
  for $t = 1, \dots, H$ **do**
    draw $A_t \sim \pi_t(\cdot \mid S_1, A_1, \dots, S_t)$ and $S_{t+1} \sim P_t(\cdot \mid S_t, A_t)$
    $w \leftarrow w + r_t(S_t, A_t, S_{t+1})$
  $W^{(i)} \leftarrow w$

Output: $\hat{V}_n(s) = \frac{1}{n}\sum_i W^{(i)}$, or the empirical measure $\hat{\nu}_n = \frac{1}{n}\sum_i \delta_{W^{(i)}}$

Proof. The first statement is the strong law of large numbers. The second follows from the linearity of expectation and the independence of the X_i: $\mathbb{V}\mathrm{ar}[\sum_i X_i] = \sum_i \mathbb{V}\mathrm{ar}[X_i] = n\sigma^2$. ∎

For the return of an MDP with rewards in $[0, R_{\max}]$, the variable W_H takes values in $[0, HR_{\max}]$, so that all moments are finite and $\sigma^2 \le (HR_{\max})^2/4$ (Exercise 2.1). The estimator is unbiased and its standard deviation decreases as $n^{-1/2}$: dividing the error by 10 requires 100 times more runs. This is the price, and the strength, of Monte Carlo methods: slow but universal.

2.2 Off-policy evaluation

In many applications the trajectories at hand were not generated by the policy π one wants to evaluate, but by another one: the recommendation system currently in production, the physicians who treated the patients in the database, the previous version of a chatbot. The trajectory law (1.1) shows that the ratio of the laws under two policies involves the policies only, not the kernels—and this makes it possible to evaluate π from the data of a *behavior policy* π_{b}. We treat this question first because it motivates much of what follows: the variance issue it raises is quantified with the tools of the next sections.

Proposition 2.2 (Importance sampling). Let $\pi_{\mathrm{b}}, \pi \in \Pi_{\mathrm{H}}$ be such that $\pi_t(a \mid h_t) > 0$ implies $\pi_{\mathrm{b},t}(a \mid h_t) > 0$ for all t, h_t, a. Define the *importance weight* of a trajectory as

$$w_H \coloneqq \prod_{t=1}^{H} \frac{\pi_t(A_t \mid H_t)}{\pi_{\mathrm{b},t}(A_t \mid H_t)},$$

which is well defined $\mathbb{P}_\mu^{\pi_{\mathrm{b}}}$-almost surely. Then for every function g on Ω_H, $\mathbb{E}_\mu^{\pi_{\mathrm{b}}}[w_H\, g] = \mathbb{E}_\mu^{\pi}[g]$. In particular $V^\pi(\mu) = \mathbb{E}_\mu^{\pi_{\mathrm{b}}}[w_H W_H]$, and if $(w^{(i)}, W^{(i)})_{i \le n}$ are computed from n independent trajectories generated under π_{b}, the estimator $\frac{1}{n}\sum_{i\le n} w^{(i)} W^{(i)}$ of $V^\pi(\mu)$ is unbiased, and so is the weighted empirical measure $\frac{1}{n}\sum_{i\le n} w^{(i)}\delta_{W^{(i)}}$ as an estimator of $\nu^\pi(\mu)$—a signed measure of random total mass $\frac{1}{n}\sum_i w^{(i)}$, not a probability measure.

Proof. By (1.1), for every $\omega \in \Omega_H$ with $\mathbb{P}_\mu^{\pi_{\mathrm{b}}}(\{\omega\}) > 0$ the factors $\mu(s_1)$ and $P_t(s_{t+1} \mid s_t, a_t)$ cancel in the ratio and $\mathbb{P}_\mu^{\pi}(\{\omega\}) = w_H(\omega)\mathbb{P}_\mu^{\pi_{\mathrm{b}}}(\{\omega\})$. If $\mathbb{P}_\mu^{\pi_{\mathrm{b}}}(\{\omega\}) = 0$, one of the factors $\mu(s_1)$, $\pi_{\mathrm{b},t}(a_t \mid h_t)$ or $P_t(s_{t+1} \mid s_t, a_t)$ vanishes, and then $\mathbb{P}_\mu^{\pi}(\{\omega\}) = 0$ as well by the support condition; such ω contribute to neither side. Hence $\mathbb{E}_\mu^{\pi}[g] = \sum_\omega g(\omega) w_H(\omega)\mathbb{P}_\mu^{\pi_{\mathrm{b}}}(\{\omega\}) = \mathbb{E}_\mu^{\pi_{\mathrm{b}}}[w_H g]$. ∎

The estimator requires no knowledge of the kernels: only the probabilities with which the behavior policy chose its actions must be recorded, which is why online services log them. Its weakness is its variance. The weight w_H has mean 1 under π_{b}, but being a product of H factors its variance can grow

exponentially with the horizon: if π is deterministic and π_b chooses uniformly among A actions, w_H equals A^H on the trajectories that follow π and 0 on the others, so that $\mathbb{V}\mathrm{ar}_{\pi_b}[w_H] = A^H - 1$. Variants exploiting the additive structure of the return (Exercise 2.9) and the dynamic programming methods of Chapter 10 mitigate this "curse of the horizon" (Liu et al., 2018), which is a central difficulty of evaluation from logged data.

> **Remark 2.3 (Variance reduction).** Importance sampling is also a variance-*reduction* device when the behavior policy is chosen on purpose: sampling more often the trajectories that contribute most to the quantity of interest (rare failures, high rewards) and reweighting by w_H can reduce the variance of the estimate below that of direct simulation of π. Together with *control variates*—subtract from $W^{(i)}$ a correlated quantity of known mean, such as the return of a policy whose value is known—these classical Monte Carlo techniques improve the constants of all the bounds of this chapter; we shall not pursue them.

2.3 Asymptotic confidence intervals

The central limit theorem describes the fluctuations of $\bar{X}_n$ around m.

> **Theorem 2.4 (Central limit theorem).** If $0 < \sigma^2 < \infty$, then $\sqrt{n}\,(\bar{X}_n - m)/\sigma$ converges in distribution to the standard Gaussian law $\mathcal{N}(0,1)$. Moreover, setting $\hat{\sigma}_n^2 \coloneqq \frac{1}{n-1}\sum_{i\le n}(X_i - \bar{X}_n)^2$, one has $\hat{\sigma}_n^2 \to \sigma^2$ almost surely and $\sqrt{n}\,(\bar{X}_n - m)/\hat{\sigma}_n$ converges in distribution to $\mathcal{N}(0,1)$.

Proof. The first claim is the classical central limit theorem. Expanding the square, $\hat{\sigma}_n^2 = \frac{n}{n-1}\big(\frac{1}{n}\sum_i X_i^2 - \bar{X}_n^2\big) \to \mathbb{E}[X_1^2] - m^2 = \sigma^2$ almost surely by the strong law of large numbers. The last claim follows from Slutsky's lemma: if Y_n converges in distribution to Y and Z_n converges in probability to a constant $c \neq 0$, then Y_n/Z_n converges in distribution to Y/c. ∎

Denoting by $z_{1-\delta/2}$ the quantile of order $1-\delta/2$ of $\mathcal{N}(0,1)$ ($z_{0.975} \approx 1.96$), we obtain the *asymptotic confidence interval*

$$\mathbb{P}\Big(m \in \Big[\bar{X}_n - z_{1-\delta/2}\frac{\hat{\sigma}_n}{\sqrt{n}},\ \bar{X}_n + z_{1-\delta/2}\frac{\hat{\sigma}_n}{\sqrt{n}}\Big]\Big) \xrightarrow[n\to\infty]{} 1-\delta. \tag{2.1}$$

It is the most commonly reported measure of uncertainty for Monte Carlo estimates, and it is sharp: the Gaussian approximation is usually excellent for moderate n. Its weakness is that the statement is only asymptotic; nothing is said about a given, finite n.

The same tool extends to smooth functionals of the mean of a transformed variable.

> **Theorem 2.5 (Delta method).** Let Y_n be random vectors in $\mathbb{R}^d$ such that $\sqrt{n}(Y_n - y)$ converges in distribution to $\mathcal{N}(0,\Sigma)$, and let $g : \mathbb{R}^d \to \mathbb{R}$ be differentiable at y. Then $\sqrt{n}\big(g(Y_n) - g(y)\big)$ converges in distribution to $\mathcal{N}\big(0, \nabla g(y)^\top \Sigma \nabla g(y)\big)$.

The proof is a first-order Taylor expansion, see van der Vaart (1998, Chap. 3). We shall use it in Section 2.5 for functionals of the form $g\big(\frac{1}{n}\sum_i f(X_i)\big)$.

2.4 Non-asymptotic guarantees

We now look for bounds valid for every n: given a precision $\varepsilon > 0$ and a confidence level $1-\delta$, how many runs guarantee $|\bar{X}_n - m| \le \varepsilon$ with probability at least $1-\delta$? The answer depends on what is assumed about the law of X_1.

2.4.1 Chebyshev's inequality

Proposition 2.6 (Markov and Chebyshev inequalities). For a nonnegative random variable Y and $u > 0$, $\mathbb{P}(Y \geq u) \leq \mathbb{E}[Y]/u$. Consequently, if $\sigma^2 < \infty$,

$$\mathbb{P}\big(|\bar{X}_n - m| \geq \varepsilon\big) \;\leq\; \frac{\sigma^2}{n\varepsilon^2} \qquad (\varepsilon > 0).$$

Proof. $u\,\mathbb{1}\{Y \geq u\} \leq Y$, and taking expectations gives Markov's inequality. Apply it to $Y = (\bar{X}_n - m)^2$ and $u = \varepsilon^2$, using $\mathbb{V}\mathrm{ar}[\bar{X}_n] = \sigma^2/n$. ∎

Chebyshev's inequality only requires a finite variance, but the resulting sample size $n \geq \sigma^2/(\delta\varepsilon^2)$ grows like $1/\delta$: guaranteeing a confidence of 99.9% costs five hundred times more than 50%. The Gaussian approximation (2.1) suggests that the true dependence should be logarithmic in $1/\delta$. Making this rigorous is the purpose of the Chernoff method.

2.4.2 The Chernoff method

Definition 2.7 (Cumulant generating function). The *cumulant generating function* of a random variable X (or of its law ν) is

$$\Lambda_X(\beta) \;\coloneqq\; \ln \mathbb{E}\big[e^{\beta X}\big] \;\in\; (-\infty, +\infty], \qquad \beta \in \mathbb{R}.$$

If X is bounded, Λ_X is finite everywhere. In general, its domain $\{\beta : \Lambda_X(\beta) < \infty\}$ is an interval containing 0, possibly reduced to $\{0\}$.

Proposition 2.8 (Chernoff bound). For every random variable X, every $u \in \mathbb{R}$ and every $\beta > 0$,

$$\mathbb{P}(X \geq u) \;\leq\; \exp\big(\Lambda_X(\beta) - \beta u\big).$$

Consequently $\mathbb{P}(X \geq u) \leq \exp\big(-\Lambda_X^*(u)\big)$ for every $u \geq \mathbb{E}[X]$, where $\Lambda_X^*(u) \coloneqq \sup_{\beta \in \mathbb{R}} \big(\beta u - \Lambda_X(\beta)\big)$ is the Legendre transform of Λ_X, also called its Cramér transform.

Proof. For $\beta > 0$ the map $x \mapsto e^{\beta x}$ is increasing, so $\{X \geq u\} = \{e^{\beta X} \geq e^{\beta u}\}$ and Markov's inequality gives $\mathbb{P}(X \geq u) \leq e^{-\beta u}\mathbb{E}[e^{\beta X}]$. Optimizing over $\beta > 0$ yields $\mathbb{P}(X \geq u) \leq \exp(-\sup_{\beta > 0}(\beta u - \Lambda_X(\beta)))$. For $u \geq \mathbb{E}[X]$ the supremum over $\beta > 0$ coincides with the supremum over $\mathbb{R}$: indeed, by Jensen's inequality $\Lambda_X(\beta) \geq \beta\mathbb{E}[X]$, so that for $\beta \leq 0$, $\beta u - \Lambda_X(\beta) \leq \beta(u - \mathbb{E}[X]) \leq 0$, which is the value of the objective at $\beta = 0$. ∎

The Chernoff method thus reduces a tail bound to a bound on exponential moments. Its power comes from the fact that exponential moments of a sum of independent variables factorize, so that Λ is additive: if X and Y are independent, $\Lambda_{X+Y} = \Lambda_X + \Lambda_Y$. It remains to bound the cumulant generating function of a single bounded variable.

Lemma 2.9 (Hoeffding's lemma). If $a \leq X \leq b$ almost surely, then for every $\beta \in \mathbb{R}$,

$$\Lambda_{X - \mathbb{E}[X]}(\beta) \;=\; \ln \mathbb{E}\big[e^{\beta(X - \mathbb{E}[X])}\big] \;\leq\; \frac{\beta^2 (b-a)^2}{8}.$$

Proof. If $a = b$ there is nothing to prove. Otherwise, replacing X by $X - \mathbb{E}[X]$ we may assume $\mathbb{E}[X] = 0$, with $a \le 0 \le b$. By convexity of the exponential, for $x \in [a,b]$, $e^{\beta x} \le \frac{b-x}{b-a}e^{\beta a} + \frac{x-a}{b-a}e^{\beta b}$. Taking expectations and using $\mathbb{E}[X] = 0$,

$$\mathbb{E}\big[e^{\beta X}\big] \;\le\; \frac{b}{b-a}e^{\beta a} - \frac{a}{b-a}e^{\beta b} \;=\; e^{\phi(\beta(b-a))}, \qquad \phi(v) \coloneqq -pv + \ln\big(1-p+pe^{v}\big), \quad p \coloneqq \frac{-a}{b-a} \in [0,1].$$

One checks that $\phi(0) = 0$, $\phi'(0) = 0$ and $\phi''(v) = \frac{pe^v(1-p)}{(1-p+pe^v)^2} \le \frac{1}{4}$, the latter because $xy \le (x+y)^2/4$ with $x = 1-p$ and $y = pe^v$. By Taylor's formula, $\phi(v) \le v^2/8$. ∎

Theorem 2.10 (Hoeffding's inequality). Let $X_1, \dots, X_n$ be independent with $a_i \le X_i \le b_i$ almost surely. For every $u > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big) \ge u\Big) \;\le\; \exp\Big(-\frac{2u^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\Big).$$

In particular, if the X_i are i.i.d. with values in $[a,b]$, for every $\varepsilon > 0$,

$$\mathbb{P}\big(|\bar{X}_n - m| \ge \varepsilon\big) \;\le\; 2\exp\Big(-\frac{2n\varepsilon^2}{(b-a)^2}\Big).$$

Proof. Let $Y = \sum_i (X_i - \mathbb{E}[X_i])$ and $c = \sum_i (b_i - a_i)^2$. By independence and Lemma 2.9, $\Lambda_Y(\beta) = \sum_i \Lambda_{X_i - \mathbb{E}X_i}(\beta) \le \beta^2 c/8$. The Chernoff bound gives $\mathbb{P}(Y \ge u) \le \exp(\beta^2 c/8 - \beta u)$ for all $\beta > 0$, and the choice $\beta = 4u/c$ yields $\exp(-2u^2/c)$. For the second statement, apply the first to $\pm(X_i - m)$ with $u = n\varepsilon$ and $c = n(b-a)^2$, and add the two bounds. ∎

The route through Lemma 2.9 is the standard one. Exercise 2.3 gives a second proof of the i.i.d. case, which goes through the Bernoulli law: for $\mathrm{Ber}(p)$ variables the Chernoff bound is explicit, $\mathbb{P}(\bar{X}_n \ge u) \le e^{-n\,\mathrm{kl}(u,p)}$ with $\mathrm{kl}(u,p)$ the Kullback–Leibler divergence between the Bernoulli laws of parameters u and p; a variable with values in $[0,1]$ has exponential moments at most those of the Bernoulli variable with the same mean, so the bound extends to it; and the elementary inequality $\mathrm{kl}(u,p) \ge 2(u-p)^2$ turns the divergence into Hoeffding's exponent. The two proofs are dual to each other—the last inequality is the Legendre transform of the bound $\beta^2/8$ of Lemma 2.9—but the second one keeps the sharper intermediate bound $e^{-n\,\mathrm{kl}(u,m)}$, which is the form we shall need in Theorem 2.28 and again in Chapter 13.

Corollary 2.11 (Sample complexity of Monte Carlo evaluation). If the rewards take values in $[0, R_{\max}]$, then for every $\varepsilon > 0$ and $\delta \in (0,1)$, the Monte Carlo estimator with

$$n \;\ge\; \frac{(HR_{\max})^2}{2\varepsilon^2}\ln\frac{2}{\delta}$$

runs satisfies $|\hat{V}_n(s) - V^{\pi}(s)| \le \varepsilon$ with probability at least $1-\delta$.

Proof. Apply Theorem 2.10 to the returns, which take values in $[0, HR_{\max}]$. ∎

The dependence in δ is now logarithmic, as announced. For the windy hike (simplest version), where $W_H \in \{0,1\}$, estimating the success probability within $\varepsilon = 0.01$ with confidence 99% requires $n \ge \ln(200)/(2\cdot 10^{-4}) \approx 26\,500$ runs.

Remark 2.12 (The curse of the horizon (Liu et al., 2018), and Bernstein's inequality). The bound of Corollary 2.11 grows like H^2, because the range of the return grows like H. When the rewards along a trajectory are weakly dependent, the variance of W_H is rather of order H, and Hoeffding's inequality is pessimistic. *Bernstein's inequality* takes the variance into account: if $|X_i - m| \le c$ almost surely and $\mathbb{V}\mathrm{ar}[X_i] = \sigma^2$, then

$$\mathbb{P}\big(\bar{X}_n - m \ge \varepsilon\big) \;\le\; \exp\Big(-\frac{n\varepsilon^2}{2\sigma^2 + 2c\varepsilon/3}\Big).$$

It is also proved by the Chernoff method, with a sharper bound on Λ; see Boucheron et al. (2013, Chap. 2). For small ε the exponent is $\approx n\varepsilon^2/(2\sigma^2)$, matching the Gaussian approximation.

2.5 Exponential moments and β-means

The Chernoff method suggests that the exponential moments $\mathbb{E}[e^{\beta X}]$ are the right quantities to describe the tails of a random variable. We now study them for their own sake. They will reappear in Chapter 4 as the statistics of the return that dynamic programming can propagate, in Chapter 5 as the utilities that it can optimize, and in Chapter 6 as the building blocks of a coherent theory of risk.

Proposition 2.13 (Properties of the cumulant generating function). Let X be such that Λ_X is finite on an open interval $J \ni 0$ (for instance, X bounded and $J = \mathbb{R}$). Then Λ_X is infinitely differentiable and convex on J, with $\Lambda_X(0) = 0$,

$$\Lambda_X'(\beta) = \mathbb{E}_\beta[X], \qquad \Lambda_X''(\beta) = \mathbb{V}\mathrm{ar}_\beta[X],$$

where $\mathbb{E}_\beta$ and $\mathbb{V}\mathrm{ar}_\beta$ refer to the *tilted law* $\mathbb{P}_\beta(\mathrm{d}x) \coloneqq e^{\beta x - \Lambda_X(\beta)}\,\mathbb{P}_X(\mathrm{d}x)$. In particular $\Lambda_X'(0) = \mathbb{E}[X]$ and $\Lambda_X''(0) = \mathbb{V}\mathrm{ar}[X]$. The convexity is strict unless X is almost surely constant.

Proof. For β in a compact subinterval of J, the derivatives $x^k e^{\beta x}$ of the integrand are dominated by an integrable function, so that differentiation under the expectation is legitimate. Writing $M(\beta) = \mathbb{E}[e^{\beta X}]$, we get $\Lambda_X' = M'/M = \mathbb{E}[Xe^{\beta X}]/M(\beta) = \mathbb{E}_\beta[X]$ and $\Lambda_X'' = M''/M - (M'/M)^2 = \mathbb{E}_\beta[X^2] - \mathbb{E}_\beta[X]^2 \ge 0$, with equality only if X is $\mathbb{P}_\beta$-almost surely constant, that is, $\mathbb{P}$-almost surely constant since the two laws are equivalent. ∎

Definition 2.14 (β-mean). Let X be a random variable and $\beta \in \mathbb{R}$ such that $\mathbb{E}[e^{\beta X}] < \infty$. The *$\beta$-mean* (or *entropic mean*) of X is

$$U_\beta(X) \;\coloneqq\; \frac{1}{\beta}\ln \mathbb{E}\big[e^{\beta X}\big] \;=\; \frac{\Lambda_X(\beta)}{\beta} \quad (\beta \neq 0), \qquad U_0(X) \coloneqq \mathbb{E}[X].$$

For a law ν, $U_\beta(\nu) \coloneqq \frac{1}{\beta}\ln \int e^{\beta x}\,\nu(\mathrm{d}x)$ denotes the β-mean of any $X \sim \nu$.

Proposition 2.15 (Properties of β-means). Let X be a random variable with Λ_X finite on an open interval $J \ni 0$.

(i) *Monotonicity in β.* The map $\beta \mapsto U_\beta(X)$ is continuous and nondecreasing on J, and $U_\beta(X) \to \mathbb{E}[X]$ as $\beta \to 0$. In particular $U_\beta(X) \ge \mathbb{E}[X]$ for $\beta > 0$ and $U_\beta(X) \le \mathbb{E}[X]$ for $\beta < 0$. It is strictly increasing unless X is almost surely constant.

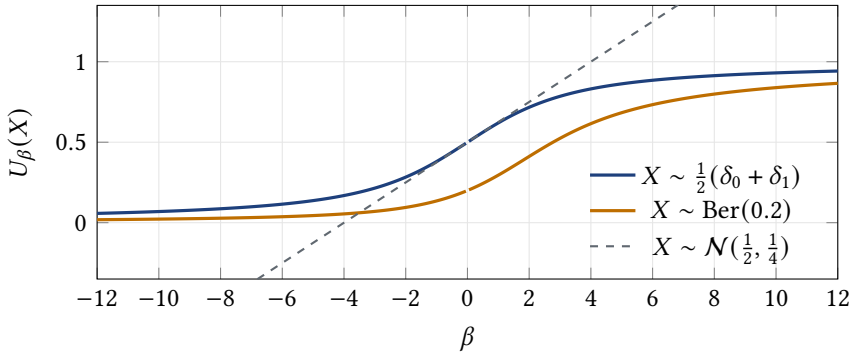


Figure 2.1. The β-mean as a function of β for three laws. For bounded laws, $U_\beta(X)$ increases from ess inf X to ess sup X; for the Gaussian law with the same mean and variance as $\frac{1}{2}(\delta_0 + \delta_1)$, it is the line $m + \beta\sigma^2/2$, which agrees with the two-point law near $\beta = 0$ and departs from it in the tails.

(ii) *Bounds and limits.* If $a \le X \le b$ almost surely, then $a \le U_\beta(X) \le b$ for all β, $U_\beta(X) \to \operatorname{ess\,sup} X$ as $\beta \to +\infty$ and $U_\beta(X) \to \operatorname{ess\,inf} X$ as $\beta \to -\infty$.

(iii) *Translation and scaling.* $U_\beta(X+c) = U_\beta(X) + c$ for $c \in \mathbb{R}$, and $U_\beta(\lambda X) = \lambda\, U_{\lambda\beta}(X)$ for $\lambda > 0$.

(iv) *Monotonicity in X.* If $X \le Y$ almost surely then $U_\beta(X) \le U_\beta(Y)$.

(v) *Additivity on independent sums.* If X and Y are independent then $U_\beta(X+Y) = U_\beta(X) + U_\beta(Y)$.

(vi) *Examples.* If $X \sim \mathcal{N}(m, \sigma^2)$ then $U_\beta(X) = m + \beta\sigma^2/2$. If $X \sim \mathrm{Ber}(p)$ then $U_\beta(X) = \frac{1}{\beta}\ln(1 - p + pe^\beta)$.

Proof. (i) For $\beta \neq 0$, $U_\beta(X) = (\Lambda_X(\beta) - \Lambda_X(0))/(\beta - 0)$ is the slope of the chord of the convex function Λ_X between 0 and β; the slope of a chord of a convex function with one fixed endpoint is a nondecreasing function of the other endpoint. Continuity away from 0 is clear, and $U_\beta(X) \to \Lambda_X'(0) = \mathbb{E}[X]$ as $\beta \to 0$ by Proposition 2.13. Strict monotonicity follows from strict convexity. (ii) For $\beta > 0$, $e^{\beta a} \le \mathbb{E}[e^{\beta X}] \le e^{\beta b}$; take logarithms and divide by β. For $\beta < 0$ the inequalities between exponentials are reversed, and dividing by $\beta < 0$ reverses them again. For the limit, let $M = \operatorname{ess\,sup} X$ and $\varepsilon > 0$: since $\mathbb{P}(X > M - \varepsilon) > 0$, $\mathbb{E}[e^{\beta X}] \ge e^{\beta(M-\varepsilon)}\mathbb{P}(X > M - \varepsilon)$, hence $U_\beta(X) \ge M - \varepsilon + \frac{1}{\beta}\ln \mathbb{P}(X > M - \varepsilon) \to M - \varepsilon$ as $\beta \to +\infty$. The case $\beta \to -\infty$ is symmetric. (iii) $\mathbb{E}[e^{\beta(X+c)}] = e^{\beta c}\mathbb{E}[e^{\beta X}]$ and $\frac{1}{\beta}\ln \mathbb{E}[e^{\beta\lambda X}] = \lambda \frac{1}{\lambda\beta}\ln \mathbb{E}[e^{(\lambda\beta)X}]$. (iv) The map $x \mapsto e^{\beta x}$ is increasing for $\beta > 0$ and decreasing for $\beta < 0$; in both cases $\frac{1}{\beta}\ln \mathbb{E}[e^{\beta \cdot}]$ is monotone nondecreasing. (v) Independence gives $\mathbb{E}[e^{\beta(X+Y)}] = \mathbb{E}[e^{\beta X}]\mathbb{E}[e^{\beta Y}]$. (vi) Direct computations: $\mathbb{E}[e^{\beta X}] = e^{\beta m + \beta^2\sigma^2/2}$ for the Gaussian law. ∎

Property (iii) shows that U_β is *not* positively homogeneous unless $\beta = 0$: the β-mean of $2X$ is $2U_{2\beta}(X)$, which exceeds $2U_\beta(X)$ for $\beta > 0$ and a nonconstant X. Property (v) is the counterpart, for the functionals U_β, of the additivity of the expectation on sums of independent variables; it will be the key to the Bellman equations for exponential moments in Chapter 4. Figure 2.1 illustrates the monotonicity and the limits of (i)–(ii).

Remark 2.16 (β-means and the mean–variance criterion). By Proposition 2.13, $\Lambda_X(\beta) = \beta\mathbb{E}[X] + \frac{\beta^2}{2}\mathbb{V}\mathrm{ar}[X] + O(\beta^3)$, hence

$$U_\beta(X) \;=\; \mathbb{E}[X] + \frac{\beta}{2}\mathbb{V}\mathrm{ar}[X] + O(\beta^2) \qquad (\beta \to 0),$$

with equality (no remainder) for Gaussian laws. For small $|\beta|$ the β-mean is thus a mean–variance criterion in the sense of Markowitz, rewarding variability when $\beta > 0$ and penalizing it when $\beta < 0$; for large $|\beta|$ it forgets the bulk of the distribution and only looks at its extremes. Contrary to $\mathbb{E}[X] - \lambda\mathbb{V}\mathrm{ar}[X]$, however, U_β is monotone in X (property (iv)): a sure improvement of the outcome never decreases it. This is one reason for preferring it as a risk-sensitive criterion, as we shall see in Chapter 6.

The Chernoff bound can be rephrased in terms of β-means, which gives them a first interpretation: $U_\beta(X)$ is the level at which the Chernoff bound at rate β becomes nontrivial, and beyond which it decays at the exponential rate β.

Corollary 2.17 (Chernoff bound in terms of β-means). For every $u \in \mathbb{R}$,

$$\mathbb{P}(X \geq u) \leq \exp\big(-\beta\,(u - U_\beta(X))\big) \quad (\beta > 0), \qquad \mathbb{P}(X \leq u) \leq \exp\big(-|\beta|\,(U_\beta(X) - u)\big) \quad (\beta < 0).$$

Proof. The first inequality is Proposition 2.8 with $\Lambda_X(\beta) = \beta U_\beta(X)$. For $\beta < 0$, $\{X \leq u\} = \{e^{\beta X} \geq e^{\beta u}\}$ and Markov's inequality gives $\mathbb{P}(X \leq u) \leq e^{-\beta u}\mathbb{E}[e^{\beta X}] = \exp(\beta(U_\beta(X) - u))$. ∎

Equivalently, for $\beta < 0$ and $\delta \in (0,1)$: with probability at least $1 - \delta$, $X \geq U_\beta(X) - \frac{1}{|\beta|}\ln\frac{1}{\delta}$. Optimizing this lower confidence bound over $\beta < 0$ will define, in Chapter 6, the entropic value at risk.

Example 2.18 (Delta method for the plug-in β-mean). The plug-in estimator of $U_\beta(\nu)$ is

$$\hat{U}_{\beta,n} = \frac{1}{\beta}\ln\Big(\frac{1}{n}\sum_{i\leq n} e^{\beta X_i}\Big).$$

Applying the central limit theorem to $Y_i = e^{\beta X_i}$ and Theorem 2.5 with $g(y) = \frac{1}{\beta}\ln y$, we get that $\sqrt{n}\,(\hat{U}_{\beta,n} - U_\beta(\nu))$ converges in distribution to $\mathcal{N}\big(0, \mathbb{V}\mathrm{ar}[e^{\beta X_1}]/(\beta^2\,\mathbb{E}[e^{\beta X_1}]^2)\big)$. A non-asymptotic counterpart is given in Proposition 2.27.

2.6 Sequential estimation and stochastic gradient descent

The Monte Carlo estimator can be computed on the fly, without storing the samples: since $\bar{X}_t = \frac{t-1}{t}\bar{X}_{t-1} + \frac{1}{t}X_t$,

$$\bar{X}_t \;=\; \bar{X}_{t-1} + \alpha_t\big(X_t - \bar{X}_{t-1}\big), \qquad \alpha_t = \frac{1}{t}. \tag{2.2}$$

The new estimate is the old one, corrected by a fraction α_t of the *prediction error* $X_t - \bar{X}_{t-1}$. This innocuous rewriting has far-reaching consequences. First, other choices of the step sizes $\alpha_t \in (0,1]$ define other estimators: a constant α gives an exponentially weighted average $\sum_{k\leq t}\alpha(1-\alpha)^{t-k}X_k$ that forgets the distant past and tracks a slowly varying mean; the conditions $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$ characterize, as we shall see in Chapter 10, the step sizes for which such recursions converge. Second, (2.2) is an instance of stochastic gradient descent: the mean m minimizes $F(u) = \frac{1}{2}\mathbb{E}[(X-u)^2]$, whose gradient is $F'(u) = u - m = \mathbb{E}[u - X]$, and $u - X_t$ is an unbiased estimate of it. Temporal-difference learning (Chapter 10) will be the same idea applied to value functions.

Theorem 2.19 (Stochastic gradient descent for convex functions). Let $\mathcal{X} \subset \mathbb{R}^d$ be closed and convex, $F : \mathcal{X} \to \mathbb{R}$ be convex with a minimizer $u^\star \in \mathcal{X}$, and $\Pi_{\mathcal{X}}$ denote the Euclidean

Algorithm 2.2: Projected stochastic gradient descent

Input: closed convex set $\mathcal{X} \subset \mathbb{R}^d$, initial point $u_1 \in \mathcal{X}$, step size $\eta > 0$, number of steps T

for $t = 1, \dots, T$ **do**
  observe a stochastic subgradient G_t of F at u_t
  $u_{t+1} \leftarrow \Pi_{\mathcal{X}}(u_t - \eta\, G_t)$

Output: the average $\bar{u}_T = \frac{1}{T}\sum_{t=1}^{T} u_t$

projection onto $\mathcal{X}$. Let (u_t) be generated by Algorithm 2.2, where $u_1 \in \mathcal{X}$ satisfies $\|u_1 - u^\star\| \le R$ and where, conditionally on the past $\mathfrak{F}_{t-1} = \sigma(u_1, G_1, \dots, G_{t-1})$, the random vector G_t satisfies $\mathbb{E}[G_t \mid \mathfrak{F}_{t-1}] \in \partial F(u_t)$ and $\|G_t\| \le L$ almost surely. Then

$$\mathbb{E}\big[F(\bar{u}_T)\big] - F(u^\star) \;\le\; \frac{R^2}{2\eta T} + \frac{\eta L^2}{2}, \qquad \text{and} \qquad \mathbb{E}\big[F(\bar{u}_T)\big] - F(u^\star) \;\le\; \frac{RL}{\sqrt{T}} \quad \text{for } \eta = \frac{R}{L\sqrt{T}}.$$

Proof. Since $u^\star \in \mathcal{X}$ and the projection onto a closed convex set is 1-Lipschitz,

$$\|u_{t+1} - u^\star\|^2 \le \|u_t - \eta G_t - u^\star\|^2 = \|u_t - u^\star\|^2 - 2\eta\langle G_t, u_t - u^\star\rangle + \eta^2\|G_t\|^2.$$

Let $g_t = \mathbb{E}[G_t \mid \mathfrak{F}_{t-1}] \in \partial F(u_t)$. As u_t is $\mathfrak{F}_{t-1}$-measurable, $\mathbb{E}[\langle G_t, u_t - u^\star\rangle \mid \mathfrak{F}_{t-1}] = \langle g_t, u_t - u^\star\rangle \ge F(u_t) - F(u^\star)$ by the subgradient inequality. Taking expectations and rearranging,

$$\mathbb{E}[F(u_t) - F(u^\star)] \le \frac{\mathbb{E}\|u_t - u^\star\|^2 - \mathbb{E}\|u_{t+1} - u^\star\|^2}{2\eta} + \frac{\eta L^2}{2}.$$

Summing over $t = 1, \dots, T$, the first terms telescope to at most $R^2/(2\eta)$. Dividing by T and using the convexity of F (Jensen's inequality) for $F(\bar{u}_T) \le \frac{1}{T}\sum_t F(u_t)$ gives the first bound; the second follows by the choice of η. ∎

The rate $T^{-1/2}$ invites comparison with the Monte Carlo rate, but the two do not measure the same thing: Theorem 2.19 controls the gap in *value*, $F(\bar{u}_T) - F(u^\star)$, where Monte Carlo controls the error on the *argument*. For the mean the two are related by a square, since $F(u) = \frac{1}{2}\mathbb{E}[(X-u)^2]$ satisfies $F(u) - F(m) = \frac{1}{2}(u-m)^2$: the Monte Carlo error $\sigma/\sqrt{T}$ is a value gap of order $1/T$, and conversely the guarantee $T^{-1/2}$ of Theorem 2.19 is only $T^{-1/4}$ on the argument. The generic bound is thus far from tight on this special case—as it must be, having used nothing beyond convexity and a bound on the gradients, while the square loss $F(u) = \frac{1}{2}\mathbb{E}[(X-u)^2]$ is as regular as a function can be: it is 1-strongly convex and its derivative is 1-Lipschitz. Both properties have names.

Definition 2.20 (Strong convexity and smoothness). A differentiable $F : \mathbb{R}^d \to \mathbb{R}$ is μ-*strongly convex* if $F(v) \ge F(u) + \langle \nabla F(u), v - u\rangle + \frac{\mu}{2}\|v - u\|^2$ for all u, v, and L-*smooth* if ∇F is L-Lipschitz, equivalently if $F(v) \le F(u) + \langle \nabla F(u), v - u\rangle + \frac{L}{2}\|v - u\|^2$ for all u, v. Comparing the two inequalities gives $\mu \le L$; the ratio $\kappa \coloneqq L/\mu \ge 1$ is the *condition number*.

Strong convexity says that the function curves upward at least quadratically, so that a small gap in value forces a small gap in the argument; smoothness says it curves upward at most quadratically, so that a gradient which is small in norm can only occur near the minimum. Together they turn the $T^{-1/2}$ of Theorem 2.19 into T^{-1}—and, this time, on both scales at once.

Theorem 2.21 (Stochastic gradient descent for strongly convex smooth functions). Let $F : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth, with minimizer $u^\star$ and condition number $\kappa = L/\mu$. Let $u_{t+1} = u_t - \eta_t G_t$, where, conditionally on $\mathfrak{F}_{t-1} = \sigma(u_1, G_1, \dots, G_{t-1})$, the random vector G_t is an unbiased estimate of the gradient with bounded variance:

$$\mathbb{E}[G_t \mid \mathfrak{F}_{t-1}] = \nabla F(u_t), \qquad \mathbb{E}\big[\|G_t - \nabla F(u_t)\|^2 \,\big|\, \mathfrak{F}_{t-1}\big] \le \sigma^2.$$

Write $a_t \coloneqq \mathbb{E}\|u_t - u^\star\|^2$. Then every step size $\eta_t \le 1/L$ satisfies

$$a_{t+1} \;\le\; (1 - \mu\eta_t)\, a_t + \eta_t^2 \sigma^2, \tag{2.3}$$

and consequently:

(i) *Constant step* $\eta \le 1/L$: $a_T \;\le\; e^{-\mu\eta(T-1)} a_1 + \dfrac{\eta\sigma^2}{\mu}$.

(ii) *Decreasing step* $\eta_t = \dfrac{2}{\mu\,(t + 2\kappa)}$:

$$a_T \;\le\; \frac{C}{T + 2\kappa}, \qquad C \coloneqq \max\left\{\frac{4\sigma^2}{\mu^2},\; (1+2\kappa)\, a_1\right\}, \qquad \text{and} \qquad \mathbb{E}[F(u_T)] - F(u^\star) \;\le\; \frac{L}{2}\, a_T.$$

Proof. Two consequences of the hypotheses, each one line. Since $\nabla F(u^\star) = 0$, smoothness applied between $u^\star$ and u, and then between u and $u - \nabla F(u)/L$, gives

$$F(u) - F(u^\star) \le \frac{L}{2}\|u - u^\star\|^2 \qquad \text{and} \qquad \|\nabla F(u)\|^2 \le 2L\big(F(u) - F(u^\star)\big), \tag{2.4}$$

the second because $F(u^\star) \le F\big(u - \nabla F(u)/L\big) \le F(u) - \frac{1}{L}\|\nabla F(u)\|^2 + \frac{1}{2L}\|\nabla F(u)\|^2$. Strong convexity applied between u and $u^\star$ gives, in the other direction,

$$\langle \nabla F(u), u - u^\star \rangle \;\ge\; F(u) - F(u^\star) + \frac{\mu}{2}\|u - u^\star\|^2. \tag{2.5}$$

Now expand the square, writing $\delta_t \coloneqq \mathbb{E}[F(u_t)] - F(u^\star) \ge 0$. As u_t is $\mathfrak{F}_{t-1}$-measurable and G_t is unbiased, the cross term has conditional expectation $\langle \nabla F(u_t), u_t - u^\star \rangle$, and $\mathbb{E}[\|G_t\|^2 \mid \mathfrak{F}_{t-1}] = \|\nabla F(u_t)\|^2 + \mathbb{E}[\|G_t - \nabla F(u_t)\|^2 \mid \mathfrak{F}_{t-1}]$. Taking expectations and using (2.5) on the cross term and (2.4) on the square term,

$$a_{t+1} \;\le\; a_t - 2\eta_t\Big(\delta_t + \frac{\mu}{2} a_t\Big) + \eta_t^2\big(2L\delta_t + \sigma^2\big) \;=\; (1 - \mu\eta_t)\, a_t - 2\eta_t(1 - L\eta_t)\, \delta_t + \eta_t^2\sigma^2.$$

For $\eta_t \le 1/L$ the middle term is nonpositive, which is (2.3).

(i) Iterating (2.3) with $\eta_t = \eta$ and summing the geometric series, $a_T \le (1-\mu\eta)^{T-1} a_1 + \eta^2\sigma^2 \sum_{k\ge 0}(1-\mu\eta)^k = (1-\mu\eta)^{T-1} a_1 + \eta\sigma^2/\mu$; conclude with $1 - x \le e^{-x}$ (and $\mu\eta \le \mu/L \le 1$).

(ii) The steps decrease and $\eta_1 = 2/(\mu + 2L) \le 1/L$, so (2.3) applies throughout. Write $s \coloneqq t + 2\kappa$, so that $\mu\eta_t = 2/s$ and $s \ge 1 + 2\kappa \ge 3$. We show $a_t \le C/s$ by induction on t; at $t = 1$ this is the second term in the definition of C. Assuming it at t, and using $s > 2$,

$$a_{t+1} \;\le\; \frac{s-2}{s} \cdot \frac{C}{s} + \frac{4\sigma^2}{\mu^2 s^2} \;\le\; \frac{C(s-2)}{s^2} + \frac{C}{s^2} \;=\; \frac{C(s-1)}{s^2} \;\le\; \frac{C}{s+1},$$

the middle inequality by the first term in the definition of C and the last because $(s-1)(s+1) \le s^2$. The bound on the values is the first half of (2.4). ∎

The same proof covers the *projected* recursion of Algorithm 2.2 whenever the unconstrained minimizer $u^\star$ lies in $\mathcal{X}$, since projecting onto a convex set that contains $u^\star$ can only decrease the distance to it.

The two step sizes are the two behaviors announced by (2.2). A constant step forgets the initial condition geometrically, at rate $\mu\eta$, and then hovers at a noise floor $\eta\sigma^2/\mu$: it tracks, and halving η halves the floor at the price of doubling the time constant. A step $\propto 1/t$ has no floor: past the transient $T \gtrsim \kappa$ the bound reads $a_T \le 4\sigma^2/(\mu^2 T)$, and $\mathbb{E}[F(u_T)] - F(u^\star) \le 2\kappa\sigma^2/(\mu T)$. This is the Monte Carlo rate on *both* scales—$T^{-1/2}$ on the argument and T^{-1} on the value—and it is the last iterate that converges, with no averaging needed. What the theorem costs is the two extra hypotheses: Theorem 2.19 asks only for convexity, and the pinball loss of Proposition 2.23 below has neither property as soon as the law of X has atoms—for a return supported by finitely many values it is piecewise linear.

Example 2.22 (The mean, once more). For $F(u) = \frac{1}{2}\mathbb{E}[(X-u)^2]$ and $G_t = u_t - X_t$ we have $\mu = L = \kappa = 1$, $u^\star = m$ and $\sigma^2 = \mathbb{V}\mathrm{ar}[X]$. With a constant step $\eta \le 1$ the recursion is $u_{t+1} = (1-\eta)u_t + \eta X_t$, the exponentially weighted average met after (2.2), whose variance converges to $\eta\sigma^2/(2-\eta)$: the floor $\eta\sigma^2$ of (i) is correct within a factor 2. With $\eta_t = 2/(t+2)$, (ii) gives $\mathbb{E}[(u_T - m)^2] \le 4\sigma^2/(T+2)$ once the transient is over, against the exact $\sigma^2/(T-1)$ achieved by $\alpha_t = 1/t$, which is the empirical mean: the right rate, with a constant four times too large. Neither the rate nor the constant can be had from Theorem 2.19, which gives only $T^{-1/4}$ on this scale.

What Theorem 2.19 gains, in exchange, is generality: any quantity defined as the minimizer of an expected convex loss can be estimated sequentially, with the same guarantee and no regularity beyond convexity. Here is an instance that will matter in Chapter 10.

Proposition 2.23 (Quantiles minimize the pinball loss). For $\tau \in (0,1)$ let $\ell_\tau(x) \coloneqq \tau x_+ + (1-\tau)(-x)_+ = x\,(\tau - \mathbb{1}\{x<0\})$ be the *pinball loss*. If $\mathbb{E}|X| < \infty$, the function $\Phi(u) \coloneqq \mathbb{E}[\ell_\tau(X-u)]$ is convex, with right derivative $\mathbb{P}(X \le u) - \tau$ and left derivative $\mathbb{P}(X < u) - \tau$ at every u. Its minimizers are exactly the τ-quantiles of X, i.e. the points u with $\mathbb{P}(X<u) \le \tau \le \mathbb{P}(X \le u)$. The stochastic subgradient $\mathbb{1}\{X_t \le u\} - \tau$ is bounded by $\max(\tau, 1-\tau) \le 1$.

Proof. ℓ_τ is convex and 1-Lipschitz, hence so is Φ, and differentiation under the expectation is justified by dominated convergence. For u fixed, the right derivative of $u \mapsto \tau(X-u)_+$ is $-\tau\mathbb{1}\{X > u\}$ and that of $u \mapsto (1-\tau)(u-X)_+$ is $(1-\tau)\mathbb{1}\{X \le u\}$; summing and taking expectations gives $\Phi'_+(u) = -\tau\mathbb{P}(X>u) + (1-\tau)\mathbb{P}(X \le u) = \mathbb{P}(X \le u) - \tau$. The left derivative is obtained similarly with $\mathbb{1}\{X \ge u\}$ and $\mathbb{1}\{X < u\}$. A convex function is minimized exactly where $\Phi'_-(u) \le 0 \le \Phi'_+(u)$. ∎

The resulting recursion $u_{t+1} = u_t + \eta\,(\tau - \mathbb{1}\{X_t \le u_t\})$ moves the estimate up by $\eta\tau$ when the sample falls above it and down by $\eta(1-\tau)$ otherwise, and Theorem 2.19 applies with $L = 1$. Applied to a return W_H, it estimates a quantile of the return distribution without ever storing a sample.

2.7 Estimating the distribution of the return

Monte Carlo simulation gives access to the whole empirical measure $\hat{\nu}_n = \frac{1}{n}\sum_{i \le n} \delta_{X_i}$, hence to the empirical distribution function $\hat{F}_n(x) = \frac{1}{n}\sum_{i \le n} \mathbb{1}\{X_i \le x\}$. How close is $\hat{\nu}_n$ to ν?

Theorem 2.24 (Glivenko–Cantelli; Dvoretzky–Kiefer–Wolfowitz–Massart). Let F be the distribution function of ν. Then $\sup_{x\in\mathbb{R}} |\hat F_n(x) - F(x)| \to 0$ almost surely, and for every $n \ge 1$ and $\varepsilon > 0$,

$$\mathbb{P}\Big(\sup_{x\in\mathbb{R}} \big|\hat F_n(x) - F(x)\big| > \varepsilon \Big) \;\le\; 2e^{-2n\varepsilon^2}.$$

For a fixed x, the bound is Hoeffding's inequality for the Bernoulli variables $\mathbb{1}\{X_i \le x\}$; the remarkable fact is that the uniform deviation obeys the same bound. We refer to Massart (1990) for the proof, which is delicate.

Uniform control of the distribution function translates into simultaneous confidence intervals for all quantiles.

Proposition 2.25 (Confidence intervals for quantiles). Let $\varepsilon_n = \sqrt{\ln(2/\delta)/(2n)}$. With probability at least $1-\delta$, for every $\tau \in (\varepsilon_n, 1-\varepsilon_n)$,

$$F^{-1}(\tau - \varepsilon_n) \;\le\; \hat F_n^{-1}(\tau) \;\le\; F^{-1}(\tau + \varepsilon_n),$$

where $G^{-1}(\tau) = \inf\{x : G(x) \ge \tau\}$ denotes the generalized inverse.

Proof. On the event $\{\sup_x |\hat F_n(x) - F(x)| \le \varepsilon_n\}$, which has probability at least $1-\delta$ by Theorem 2.24: if $\hat F_n(x) \ge \tau$ then $F(x) \ge \tau - \varepsilon_n$, hence $x \ge F^{-1}(\tau-\varepsilon_n)$; taking $x = \hat F_n^{-1}(\tau)$, for which $\hat F_n(x) \ge \tau$ by right-continuity, gives the lower bound. Conversely, with $x = F^{-1}(\tau+\varepsilon_n)$ we have $F(x) \ge \tau + \varepsilon_n$, so $\hat F_n(x) \ge \tau$ and $\hat F_n^{-1}(\tau) \le x$. ∎

The closeness of $\hat\nu_n$ to ν can also be measured by the Wasserstein distance $\mathcal{W}_1$, which we define and study in Appendix D; we only need here that on the real line $\mathcal{W}_1(\nu,\nu') = \int_{\mathbb{R}} |F_\nu(x) - F_{\nu'}(x)|\,dx$.

Proposition 2.26 (Wasserstein consistency of the empirical measure). If ν is supported in $[a,b]$, then $\mathbb{E}\big[\mathcal{W}_1(\hat\nu_n,\nu)\big] \le \frac{1}{\sqrt n}\int_a^b \sqrt{F(x)(1-F(x))}\,dx \le \frac{b-a}{2\sqrt n}$.

Proof. By Fubini's theorem and the Cauchy–Schwarz inequality, $\mathbb{E}\int |\hat F_n - F|\,dx = \int \mathbb{E}|\hat F_n(x) - F(x)|\,dx \le \int \sqrt{\mathbb{V}\mathrm{ar}[\hat F_n(x)]}\,dx$, and $\mathbb{V}\mathrm{ar}[\hat F_n(x)] = F(x)(1-F(x))/n \le 1/(4n)$. ∎

Plug-in estimators of functionals that are Lipschitz for $\mathcal{W}_1$—utilities with Lipschitz f, distorted means, see Corollary D.4—inherit this rate. Quantiles are more delicate: Proposition 2.25 is uniform in the *level* τ, but it bounds the distance $|\hat F_n^{-1}(\tau) - F^{-1}(\tau)|$ on the real line only where F increases at a definite rate, and no rate holds where the density vanishes. For β-means, whose sensitivity to the tails is governed by β, a direct argument is more informative.

Proposition 2.27 (Plug-in estimation of β-means). Let $X_1,\dots,X_n$ be i.i.d. with values in $[a,b]$, $\beta > 0$, and $\hat U_{\beta,n} = \frac{1}{\beta}\ln\big(\frac{1}{n}\sum_{i\le n} e^{\beta X_i}\big)$. If $n \ge 2e^{2\beta(b-a)}\ln(2/\delta)$, then with probability at least $1-\delta$,

$$\big|\hat U_{\beta,n} - U_\beta(X_1)\big| \;\le\; \frac{2}{\beta}\,e^{\beta(b-a)}\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

The same holds for $\beta < 0$ with $|\beta|$ in place of β.

Proof. Let $Y_i = e^{\beta X_i} \in [e^{\beta a}, e^{\beta b}]$, $M = \mathbb{E}[Y_1] \geq e^{\beta a}$ and $\Delta = (e^{\beta b} - e^{\beta a})\sqrt{\ln(2/\delta)/(2n)}$. By Theorem 2.10, $|\bar{Y}_n - M| \leq \Delta$ with probability at least $1 - \delta$. The condition on n ensures $\Delta \leq e^{\beta b}\sqrt{\ln(2/\delta)/(2n)} \leq e^{\beta a}/2 \leq M/2$, so that on this event $\bar{Y}_n \geq M/2$ and, the logarithm being $2/M$-Lipschitz on $[M/2, \infty)$, $|\ln \bar{Y}_n - \ln M| \leq 2\Delta/M \leq 2e^{-\beta a}\Delta \leq 2e^{\beta(b-a)}\sqrt{\ln(2/\delta)/(2n)}$. Divide by β. ∎

The bound of Proposition 2.27 is the crudest possible: it treats $e^{\beta X}$ as a bounded variable and pays the whole range $e^{\beta(b-a)}$. A sharper argument keeps the exponential structure and gives, for variables in $[0, 1]$, a bound of exactly the Chernoff form, with the accuracy ε replaced by its image under a warping that depends on β.

Theorem 2.28 (Concentration of the plug-in β-mean, Marthe et al., 2026). Let $X_1, \ldots, X_n$ be i.i.d. with values in $[0, 1]$, let $\beta \neq 0$, write $\mu_\beta = U_\beta(X_1)$ and $\hat{\mu}_{\beta,n} = \frac{1}{\beta}\ln\left(\frac{1}{n}\sum_i e^{\beta X_i}\right)$, and set

$$g_\beta(x) = \frac{e^{\beta x} - 1}{e^\beta - 1} \qquad (x \in [0, 1]).$$

Then for every $0 < \varepsilon < 1 - \mu_\beta$,

$$\begin{aligned}\mathbb{P}\big(\hat{\mu}_{\beta,n} > \mu_\beta + \varepsilon\big) &\leq \exp\Big(-n\ \mathrm{kl}\big(g_\beta(\mu_\beta + \varepsilon),\, g_\beta(\mu_\beta)\big)\Big)\\ &\leq \exp\Big(-2n\big(g_\beta(\mu_\beta + \varepsilon) - g_\beta(\mu_\beta)\big)^2\Big),\end{aligned}$$

and symmetrically for the lower deviation.

Proof. The map g_β is an increasing bijection of $[0, 1]$ onto itself, and $Y_i = g_\beta(X_i) = (e^{\beta X_i} - 1)/(e^\beta - 1)$ takes values in $[0, 1]$ with $\mathbb{E}[Y_1] = g_\beta(\mu_\beta)$, by the definition of μ_β. The event $\{\hat{\mu}_{\beta,n} > \mu_\beta + \varepsilon\}$ is, for $\beta > 0$, $\{\frac{1}{n}\sum_i e^{\beta X_i} > e^{\beta(\mu_\beta+\varepsilon)}\}$, that is, $\{\bar{Y}_n > g_\beta(\mu_\beta + \varepsilon)\}$; the Chernoff–Hoeffding bound for variables in $[0, 1]$, $\mathbb{P}(\bar{Y}_n \geq q) \leq e^{-n\,\mathrm{kl}(q,\mathbb{E}Y_1)}$ for $q > \mathbb{E}Y_1$ (Exercise 2.3(a) and (c): a variable in $[0, 1]$ of given mean has the largest exponential moments when it is Bernoulli), gives the first inequality, and Pinsker's inequality $\mathrm{kl}(q, p) \geq 2(q - p)^2$ (Exercise 2.3(b)) the second. For $\beta < 0$ the map g_β is still increasing—$g'_\beta(x) = \beta e^{\beta x}/(e^\beta - 1) > 0$, both factors changing sign together—and the equivalence reverses twice, once on dividing by $\beta < 0$ and once because $e^\beta - 1 < 0$ makes $y \mapsto (y - 1)/(e^\beta - 1)$ decreasing; the event is again $\{\bar{Y}_n > g_\beta(\mu_\beta + \varepsilon)\}$ and the same computation applies. ∎

Three readings. The bound *is* Hoeffding's, for the transformed variable: as $\beta \to 0$, $g_\beta(x) \to x$ and one recovers $\exp(-2n\varepsilon^2)$ exactly. The quantity that plays the role of the accuracy is $g_\beta(\mu_\beta + \varepsilon) - g_\beta(\mu_\beta)$, and this is where the difficulty of large rates lives: for X uniform on $[0, 1]$ it equals $(e^{\beta\varepsilon} - 1)/\beta$, which tends to 0 as $\beta \to -\infty$, so the bound on the *upper* deviation becomes vacuous; symmetrically $g_\beta(\mu_\beta) - g_\beta(\mu_\beta - \varepsilon) = (1 - e^{-\beta\varepsilon})/\beta \to 0$ as $\beta \to +\infty$ and the bound on the lower deviation becomes vacuous. Each extreme kills the tail that U_β is looking at: estimating a β-mean at a large $|\beta|$ is estimating an extreme of the distribution, and the Bernoulli domination that produced the bound is exactly what is lost there (Exercise 2.5). Finally the estimator is biased, and its bias has a sign: by Jensen's inequality applied to the concave logarithm, $\mathbb{E}[\hat{\mu}_{\beta,n}] \leq \mu_\beta$ for $\beta > 0$ and $\geq \mu_\beta$ for $\beta < 0$—the plug-in rule is optimistic about a risk-averse criterion, which is the wrong direction for a safety-critical application.

Remark 2.29 (Estimating the whole curve at once). In Chapter 5 the rate β is a dial to be turned, so what one wants is not $\hat{\mu}_\beta$ for one rate but the function $\beta \mapsto \hat{\mu}_\beta$. It comes for free. Writing $i(X, \varepsilon) = \inf_\beta\{g_\beta(\mu_\beta + \varepsilon) - g_\beta(\mu_\beta)\}$, if X takes the values 0 and 1 with positive probability then $i(X, \varepsilon) > 0$, and the Dvoretzky–Kiefer–Wolfowitz inequality (Theorem 2.24) together with

the Lipschitz dependence of $\mathbb{E}[e^{\beta X}]$ on the distribution function gives

$$\mathbb{P}\Big(\sup_{\beta\in\mathbb{R}}\big|\hat{\mu}_{\beta,n}-\mu_\beta\big|>\varepsilon\Big)\ \le\ 2\exp\big(-2n\,i(X,\varepsilon)^2\big)$$

(Marthe et al., 2026): one sample of size n estimates the entire curve at the rate at which it estimates one point, because the whole family is a Lipschitz functional of the empirical distribution function. The constant $i(X,\varepsilon)$ is distribution-dependent and can be arbitrarily small, so this is not a uniform statement over distributions; and on a window $\beta\in[B,0]$ it degrades like $|B|e^{-|B|}$, which is what makes the plug-in estimation of the entropic value at risk—a supremum over β, Chapter 6—converge slowly.

The factor $e^{\beta(b-a)}$ is the price of looking at the tails: for large $|\beta|$, the β-mean is determined by rare extreme values, which are rarely observed. The Monte Carlo approach estimates every regular functional of the return distribution at the universal rate $n^{-1/2}$, but the constants can be prohibitive.

2.8 Censored observations: the bakery revisited

We return to the bakery of Example 1.1. The newsvendor formula requires the distribution function F of the demand; the baker does not know it and must learn it from experience. The difficulty is that she does not observe the demand D_i of day i, but only the sales $Y_i = A_i \wedge D_i$: when the bakery sells out, the unmet demand is invisible. The data are *censored* by the decisions.

2.8.1 The Kaplan–Meier estimator

Suppose the demand takes integer values and let $h_j := \mathbb{P}(D=j \mid D\ge j)$ be its *hazard* at level j (defined when $\mathbb{P}(D\ge j)>0$). Since $\mathbb{P}(D\ge j+1)=\mathbb{P}(D\ge j)(1-h_j)$ and $\mathbb{P}(D\ge 0)=1$, the survival function factorizes as

$$S(u)\ :=\ \mathbb{P}(D>u)\ =\ \prod_{j=0}^{u}(1-h_j),\qquad u\in\mathbb{N}. \tag{2.6}$$

The point of this rewriting is that each hazard h_j can be estimated from the days on which the event $\{D_i=j\}$ was *observable*, namely the days with $A_i>j$: on such a day $Y_i\ge j$ if and only if $D_i\ge j$, and $Y_i=j$ if and only if $D_i=j$. Let $n_j=|\{i\le n: A_i>j,\ Y_i\ge j\}|$ be the number of days "at risk" at level j and $d_j=|\{i\le n: A_i>j,\ Y_i=j\}|$ the number of days on which the demand was exactly j among them. The *Kaplan–Meier* (or product-limit) estimator is

$$\hat{S}_n(u)\ =\ \prod_{j=0}^{u}\Big(1-\frac{d_j}{n_j}\Big),\qquad \hat{F}_n^{\mathrm{KM}}(u)=1-\hat{S}_n(u), \tag{2.7}$$

with the convention $d_j/n_j=0$ when $n_j=0$.

Proposition 2.30 (Consistency under adaptive censoring, Huh et al., 2011). Assume that the demands $D_1,D_2,\dots$ are i.i.d. with law F and that each order A_i is a function of the past observations $(A_1,Y_1,\dots,A_{i-1},Y_{i-1})$ (and possibly of independent randomness). If the numbers $n_j=n_j^{(n)}$ of days at risk tend to infinity almost surely for every $j\le u$, then $\hat{S}_n(u)\to S(u)$ almost surely.

Proof sketch. Fix j with $h_j\in(0,1)$ (if $h_j\in\{0,1\}$, then $d_j/n_j=h_j$ exactly as soon as $n_j>0$). On the event $\{A_i>j,\ D_i\ge j\}$, which is determined by A_i (a function of the past) and by D_i, the conditional

probability that $D_i = j$ is h_j, because D_i is independent of the past. Hence $d_j - h_j n_j = \sum_{i \le n} \mathbb{1}\{A_i > j,\, Y_i \ge j\}\big(\mathbb{1}\{Y_i = j\} - h_j\big)$ is a martingale with bounded increments, and the strong law of large numbers for martingales (Theorem A.4) gives $d_j/n_j \to h_j$ on $\{n_j \to \infty\}$. Take the product over $j \le u$. ∎

The Kaplan–Meier estimator was introduced in survival analysis (Kaplan and Meier, 1958), where the "demand" is a lifetime and the "order" is the duration of the follow-up; it is one of the most used estimators in statistics. Its transfer to inventory control, where the censoring level is chosen by the decision maker and is therefore *adaptive* rather than exogenous, is due to Huh et al. (2011), who prove Proposition 2.30 and build the ordering policy on top of it.

2.8.2 Learning to order without observing the demand

There is a more direct route, due to Huh and Rusmevichientong (2009): instead of estimating F and plugging it into the newsvendor formula, optimize the expected profit $\varphi(a) = \mathbb{E}[R(a)]$ by stochastic gradient ascent. The proof of Proposition 1.2 shows that φ is concave with left derivative $g - (g+\ell)F(a^-) = g - (g+\ell)\mathbb{P}(D < a)$ at a. The crucial observation is that the event $\{D_i < a\}$ is observable when $A_i = a$: it coincides with $\{Y_i < A_i\}$, "the bakery did not sell out".

Proposition 2.31 (Online newsvendor, Huh and Rusmevichientong, 2009). Assume that the demands D_i are i.i.d. with distribution function F and that each order A_i is a function of $(A_1, Y_1, \dots, A_{i-1}, Y_{i-1})$ and of independent randomness. Then $G_i := g - (g+\ell)\mathbb{1}\{Y_i < A_i\}$ satisfies $\mathbb{E}[G_i \mid A_1, Y_1, \dots, A_i] = g - (g+\ell)F(A_i^-)$, which is a supergradient of the concave function φ at A_i, and $|G_i| \le \max(g, \ell)$. Consequently, the projected stochastic gradient ascent $A_{i+1} = \Pi_{[0, a_{\max}]}(A_i + \eta G_i)$ with $\eta = a_{\max}/(\max(g,\ell)\sqrt{T})$ satisfies

$$\varphi(a^\star) - \mathbb{E}\Big[\varphi\Big(\frac{1}{T}\sum_{i=1}^{T} A_i\Big)\Big] \;\le\; \frac{a_{\max}\max(g,\ell)}{\sqrt{T}},$$

where $a^\star$ maximizes φ over $[0, a_{\max}]$ (if $a_{\max}$ exceeds the largest possible demand, $a^\star$ is the newsvendor quantity of Proposition 1.2).

Proof. Given the past and A_i, $\mathbb{1}\{Y_i < A_i\} = \mathbb{1}\{D_i < A_i\}$ has conditional expectation $\mathbb{P}(D < A_i) = F(A_i^-)$, by independence of D_i from the past. For a concave function, the left derivative is a supergradient. Apply Theorem 2.19 to $F = -\varphi$ on $\mathcal{X} = [0, a_{\max}]$, with $R = a_{\max}$ and $L = \max(g, \ell)$. ∎

The baker thus learns the optimal order without ever seeing a demand, by a rule of striking simplicity: order a bit more after a sell-out, a bit less after a day with leftovers. This is the adaptive inventory management (AIM) algorithm of Huh and Rusmevichientong (2009) for the perishable newsvendor. This is a first example of *learning by interaction*, with a guarantee that improves as $T^{-1/2}$. We shall meet the same pattern—an unbiased estimate of a gradient computed from what is actually observed—in the temporal-difference and policy-gradient methods of Chapters 10 and 12.

Distributional viewpoint

Simulation is the most general way of studying a policy: Algorithm 2.1 returns not a number but a sample of the return, hence the whole empirical measure $\hat{\nu}_n$, and every functional of $\nu^\pi(s)$ that is Lipschitz for $\mathcal{W}_1$ or for the sup-norm on distribution functions—the mean, Lipschitz utilities, β-means, distorted means—is estimated by plug-in at the rate $n^{-1/2}$ (Theorem 2.24 and Proposition 2.26), while quantiles are located at that rate in level rather than in value. The

price is in the constants: the range of the return grows with the horizon, and functionals sensitive to the tails, such as β-means with large $|\beta|$, are hard to estimate (Proposition 2.27). The Chernoff method suggests that the exponential moments $\mathbb{E}[e^{\beta W_H}]$ are the natural currency for describing tails; Chapter 4 will show that they are also the statistics that dynamic programming can compute exactly.

Hands-on

In `bakery.ipynb` (session H01), the agent facing an unknown demand law observes censored sales only. This chapter suggests two strategies: estimate the demand distribution by the Kaplan–Meier estimator (2.7)—the script `kaplanMeyer.py` implements it on simulated censored data—and plug it into the newsvendor formula; or run the online newsvendor of Proposition 2.31. Comparing their profits by Monte Carlo, with the confidence intervals of (2.1) and Theorem 2.10, is a good exercise.

Exercises

Exercise 2.1 (Variance of a bounded variable). Show that if $a \le X \le b$ almost surely then $\mathbb{V}\mathrm{ar}[X] \le (b-a)^2/4$, with equality if and only if X takes the values a and b with probability $1/2$ each. *Hint:* $\mathbb{V}\mathrm{ar}[X] \le \mathbb{E}[(X-c)^2]$ for $c = (a+b)/2$.

Exercise 2.2 (Cramér transform). Compute Λ_X and Λ_X^* for $X \sim \mathcal{N}(m, \sigma^2)$ and for $X \sim \mathrm{Ber}(p)$. Show that in the Bernoulli case $\Lambda_X^*(u) = \mathrm{kl}(u,p) \coloneqq u \ln \frac{u}{p} + (1-u)\ln\frac{1-u}{1-p}$ for $u \in [0,1]$, and compare the Chernoff bound $\mathbb{P}(\bar{X}_n \ge u) \le e^{-n\,\mathrm{kl}(u,p)}$, valid for $u \in [p,1]$, with Hoeffding's inequality. Prove that Λ_X^* is convex, nonnegative, and vanishes at $u = \mathbb{E}[X]$.

Exercise 2.3 (Hoeffding's inequality through the Bernoulli case). This exercise gives a second proof of Theorem 2.10, which bypasses Lemma 2.9 and goes through the Kullback–Leibler divergence $\mathrm{kl}(u,p)$ of Exercise 2.2. (a) Let $X_1, \dots, X_n$ be i.i.d. $\mathrm{Ber}(p)$ with $p \in (0,1)$. Prove Chernoff's inequality in its Bernoulli form: $\mathbb{P}(\bar{X}_n \ge u) \le e^{-n\,\mathrm{kl}(u,p)}$ for every $u \in [p,1)$. (b) Prove Pinsker's inequality for Bernoulli laws, $\mathrm{kl}(u,p) \ge 2(u-p)^2$ for all $u, p \in (0,1)$, directly: at fixed p, the function $u \mapsto \mathrm{kl}(u,p) - 2(u-p)^2$ vanishes together with its derivative at $u = p$, and its second derivative is nonnegative. Deduce Hoeffding's inequality for Bernoulli variables, $\mathbb{P}(\bar{X}_n \ge p + \varepsilon) \le e^{-2n\varepsilon^2}$. (c) Let X take values in $[0,1]$ with $\mathbb{E}[X] = m$. Show that $\mathbb{E}[e^{\beta X}] \le 1 - m + me^{\beta}$ for every $\beta \in \mathbb{R}$: the exponential moments of X are at most those of a Bernoulli variable with the same mean. Conclude that (a) and (b) hold, with p replaced by m, for i.i.d. variables with values in $[0,1]$, and recover the i.i.d. case of Theorem 2.10 by an affine change of variable. (d) Show that (b) is equivalent to Lemma 2.9 for Bernoulli variables: the Legendre transforms of $u \mapsto \mathrm{kl}(u,m)$ and of $u \mapsto 2(u-m)^2$ are $\beta \mapsto \ln(1-m+me^{\beta})$ and $\beta \mapsto \beta m + \beta^2/8$. Extend (c) to independent variables with different means (*hint:* $m \mapsto \ln(1-m+me^{\beta})$ is concave).

Exercise 2.4 (Scaling and additivity of β-means). Show that for $X \sim \mathcal{N}(m, \sigma^2)$ and $\lambda > 0$, $U_\beta(\lambda X) \ne \lambda U_\beta(X)$ unless $\beta = 0$ or $\lambda = 1$. Show that $\beta \mapsto \beta U_\beta(X)$ is convex. Using the Cauchy–Schwarz inequality, show that for arbitrary (possibly dependent) X and Y, $U_\beta(X+Y) \le$

$U_{2\beta}(X) + U_{2\beta}(Y)$ if $\beta > 0$, and the reverse inequality if $\beta < 0$. Finally, taking $Y = X$ and then $Y = -X$, show that for $\beta \neq 0$ the functional U_β is in general neither subadditive nor superadditive.

Exercise 2.5 (Estimating a β-mean at a large rate). Let X be uniform on $[0,1]$ and $g_\beta(x) = (e^{\beta x} - 1)/(e^\beta - 1)$ as in Theorem 2.28. (a) Compute $\mu_\beta = U_\beta(X)$ in closed form and check that $g_\beta(\mu_\beta) = \frac{1}{\beta} + \frac{1}{1-e^\beta}$. (b) Show that $g_\beta(\mu_\beta + \varepsilon) - g_\beta(\mu_\beta) = (e^{\beta\varepsilon} - 1)/\beta$ and that it tends to 0 as $\beta \to -\infty$ for every fixed ε, so that the bound of Theorem 2.28 on the upper deviation becomes vacuous; state and prove the mirror statement for the lower deviation as $\beta \to +\infty$. (c) Explain this in words by identifying $\lim_{\beta\to-\infty} \hat\mu_{\beta,n}$ and $\lim_{\beta\to-\infty} \mu_\beta$, and deduce that no concentration bound can be uniform in β for a variable with a continuous law. (d) Show that if instead $\mathbb{P}(X = 0) > 0$ and $\mathbb{P}(X = 1) > 0$, the infimum $i(X,\varepsilon)$ of Remark 2.29 is positive, and compute it for $X \sim \mathrm{Ber}(p)$.

Exercise 2.6 (Bernstein versus Hoeffding). In the windy hike of Example 1.4 (simplest version), suppose that the success probability is $V^\pi(\mathrm{S}) = 0.05$. Compare the number of runs needed to estimate it within $\varepsilon = 0.01$ with probability 0.99 according to Theorem 2.10 and to the two-sided version of Bernstein's inequality (Remark 2.12, proved as Theorem B.2 in Appendix B), with the true variance and $c = 0.95$. Comment.

Exercise 2.7 (Recursive estimation with general step sizes). Let $\hat m_t = \hat m_{t-1} + \alpha_t(X_t - \hat m_{t-1})$ with $\hat m_0 = 0$. Show that $\hat m_t = \sum_{k\le t} w_{t,k} X_k$ with weights $w_{t,k} = \alpha_k \prod_{j=k+1}^{t}(1-\alpha_j)$ summing to $1 - \prod_{j\le t}(1-\alpha_j)$. Compute the weights for $\alpha_t = 1/t$ and for $\alpha_t = \alpha$ constant. Show that (when all $\alpha_t < 1$) $\mathbb{E}[\hat m_t] \to m$ for every law of mean $m \neq 0$ if and only if $\sum_t \alpha_t = \infty$, and that $\mathbb{V}\mathrm{ar}[\hat m_t] \to 0$ if in addition $\alpha_t \to 0$.

Exercise 2.8 (Quantile tracking). Implement the recursion $u_{t+1} = u_t + \eta(\tau - \mathbb{1}\{X_t \le u_t\})$ of Proposition 2.23 on simulated returns of the bike store of Example 1.3 and compare with the empirical quantile and the confidence interval of Proposition 2.25. What happens when η is constant and the distribution of the X_t changes over time?

Exercise 2.9 (Per-decision importance sampling, after Precup et al., 2000). In the setting of Proposition 2.2, let $w_t = \prod_{k\le t} \pi_k(A_k \mid H_k)/\pi_{\mathrm{b},k}(A_k \mid H_k)$ be the importance weight up to time t. Show that $\mathbb{E}_\mu^{\pi_\mathrm{b}}[w_t R_t] = \mathbb{E}_\mu^{\pi}[R_t]$, hence that $\sum_{t\le H} w_t R_t$ is an unbiased estimator of $V^\pi(\mu)$ (*hint:* R_t depends on the trajectory up to time $t+1$ only). In the example of Section 2.2 (π deterministic, π_b uniform over A actions) with all rewards equal to 1, compute the variances of $w_H W_H$ and of $\sum_t w_t R_t$ and compare. Show that the *self-normalized* estimator $\sum_i w^{(i)} W^{(i)} / \sum_i w^{(i)}$ is biased but consistent, and that it always lies within the range of the observed returns.

Exercise 2.10 (Explore then commit). In the treatments problem of Example 1.9 with $K = 2$ treatments of means $\theta_1 > \theta_2$, the physician tries each treatment on n patients, then prescribes the one with the higher empirical mean to the remaining $H - 2n$ patients. Using Theorem 2.10 for the $2n$ independent variables $X_{2,i}$ and $-X_{1,i}$ at once, bound the probability that she commits to the wrong treatment by $\exp(-n(\theta_1 - \theta_2)^2)$, and deduce a bound on the expected total outcome lost with respect to the best treatment (the *regret*). Which n minimizes this bound when $\theta_1 - \theta_2 = \Delta$ is known? What happens when Δ is small, and why is this unsatisfactory? Chapter 13 presents

better strategies.

Exercise 2.11 (Kaplan–Meier without censoring). Show that if $A_i > D_i$ for all i (no censoring), the Kaplan–Meier estimator (2.7) coincides with the empirical survival function $\frac{1}{n}\sum_i \mathbb{1}\{D_i > u\}$.

Exercise 2.12 (Online newsvendor with a strictly convex loss). Suppose the demand has a density bounded below by $c > 0$ on $[0, a_{\max}]$. Show that $-\varphi$ is $c(g+\ell)$-strongly convex on this interval, and that stochastic gradient ascent with step sizes $\eta_i = 1/(c(g+\ell)\, i)$ achieves an expected suboptimality of order $(\ln T)/T$ (see Bubeck (2015, Chap. 6) for the corresponding theorem).

Bibliographical notes

Concentration inequalities are the subject of the monograph of Boucheron et al. (2013), where the Chernoff method, Hoeffding's and Bernstein's inequalities are treated in detail. Hoeffding's inequality is from Hoeffding (1963), whose Theorem 1 is the Kullback–Leibler form $e^{-n\,\mathrm{kl}(m+\varepsilon,m)}$ of Exercise 2.3 and Theorem 2 the form $e^{-2n\varepsilon^2}$ of Theorem 2.10; the reduction of a $[0,1]$-valued variable to the Bernoulli variable with the same mean is his as well. The Chernoff bound is from Chernoff (1952), and the Cramér transform from Cramér (1938), who used it to prove the first large deviation theorem. The β-mean is known under many names—exponential certainty equivalent, entropic risk measure, free energy—and its role in risk-sensitive control goes back to Howard and Matheson (1972); we come back to this history in Chapter 6.

Stochastic approximation was founded by Robbins and Monro (1951). The convergence rate of Theorem 2.19 for convex Lipschitz functions is due to Nemirovski and Yudin (1983); a modern account is Bubeck (2015), and Nemirovski et al. (2009) is the reference for the tuning of the step size and for the robustness of the averaged iterate. Theorem 2.21 is the textbook analysis of the strongly convex smooth case, in the form given by Bottou et al. (2018, Sect. 4.2)—the recursion (2.3) with its two regimes, a geometric transient and a noise floor proportional to the step—with the non-asymptotic constants of Moulines and Bach (2011). Without smoothness, strong convexity alone still gives $O(L^2 \ln T/(\mu T))$ for the uniformly averaged iterate and $2L^2/(\mu(T+1))$ for the average weighted by t (Lacoste-Julien et al., 2012); the last iterate is then no longer the right thing to look at. The Polyak–Ruppert average of the iterates of a constant-step recursion recovers the optimal asymptotic variance (Polyak and Juditsky, 1992). Quantile regression via the pinball loss is due to Koenker and Bassett (1978). The Glivenko–Cantelli theorem dates back to 1933; the inequality of Dvoretzky et al. (1956) was given its sharp constant by Massart (1990). The rate of convergence of empirical measures in Wasserstein distance is studied in depth by Bobkov and Ledoux (2019). The delta method and Slutsky's lemma are in van der Vaart (1998).

Section 2.8 follows the data-driven inventory literature. The Kaplan–Meier estimator is from Kaplan and Meier (1958); its use under decision-dependent censoring, and the consistency of Proposition 2.30, are due to Huh et al. (2011). The online newsvendor of Proposition 2.31 is the AIM algorithm of Huh and Rusmevichientong (2009), whose analysis of the perishable case gives the projection set, the step size and the $O(T^{-1/2})$ rate reproduced here; Exercise 2.12 is their strongly convex refinement. Exercise 2.9 is per-decision importance sampling, named and analyzed by Precup et al. (2000); its first claim is Exercise 5.13 of Sutton and Barto (2018, Sects. 5.5 and 5.9), and Exercise 2.7 is their Exercise 2.4. The "curse of the horizon" of Remark 2.12 is named by Liu et al. (2018).

Theorem 2.28 and the uniform-in-β bound of Remark 2.29 are due to Marthe et al. (2026), with the proofs and the discussion of what happens at large rates in Marthe (2026, Chap. 5); the transformation g_β that turns the estimation of a β-mean into the estimation of a Bernoulli parameter is theirs.