Chapter 11

From online learning to bandits

Chapter summary

- Beyond empirical and expected risk minimization, more complex settings can be considered.
- Online convex optimization with gradients: stochastic gradient descent still works with the regret criterion and potentially adversarial functions, with essentially the same rates. The mirror descent framework is adapted to non-Euclidean geometries.
- Zero-th order optimization: Randomization can be used to obtain a stochastic gradient from function values with an additional dimension dependence.
- Multi-armed bandits: in the regret minimization framework, to tackle exploration/exploitation trade-offs, several algorithms can be used, from simple algorithms based on alternating exploration and exploitation to more refined ones utilizing the principle of "optimism in the face of uncertainty."

In traditional stochastic optimization such as presented in Chapter 5 (e.g., Section 5.4), we observe a sequence of gradients of loss functions obtained from a pair of observations $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$:

$$F'_t(\theta_{t-1}) = \frac{\partial \ell(y_t, f_\theta(x_t))}{\partial \theta}\Big|_{\theta=\theta_{t-1}},$$

and our performance measure was

$$\mathbb{E}\big[F(\theta_t)\big] - F_*,$$

where the expectation is taken with respect to the training data, and $F(\theta) = \mathbb{E}[\ell(y_s, f_\theta(x_s))]$ is the expected test error, assuming that all (x_s, y_s) (and thus the individual loss functions $F_s(\theta) = \ell(y_s, f_\theta(x_s))$), $s = 1, \dots, t$, are independent and identically distributed, and F_* is the minimal value of F, that is, $F_* = \inf_{\theta \in \mathcal{C}} F(\theta)$, where $\mathcal{C}$ is the optimization domain.

There are several important extensions corresponding to specific applications:

- **Regret instead of final performance:** The performance criterion can take into account performance along iterations such as $\frac{1}{t}\sum_{s=1}^{t} F(\theta_{s-1})$, and not only at the last iteration, that is $F(\theta_t)$. This is important when the loss functions can be interpreted as actual financial losses incurred while learning the parameter θ (such as in applications in advertising or finance).

  Performance measures such as the *regret* can then be considered, here equal to

  $$\frac{1}{t}\sum_{s=1}^{t} F(\theta_{s-1}) - \inf_{\theta\in\mathcal{C}} F(\theta),$$

  often after taking the expectation (since θ_s is random because it depends on the past data).

  In this book, we choose to study what is often called the *normalized* regret since we divide $\sum_{s=1}^{t}\left[F(\theta_s) - \inf_{\theta\in\mathcal{C}} F(\theta)\right]$ by t. This is done to make comparisons with the usual stochastic framework easier.

- **Adversarial instead of stochastic:** The consideration of the regret criterion opens up the possibility for functions F_s to be different or sampled from different distributions, with a potentially adversarial choice that depends on the past. The regret is then $\frac{1}{t}\sum_{s=1}^{t} F_s(\theta_{s-1}) - \inf_{\theta\in\mathcal{C}} \frac{1}{t}\sum_{s=1}^{k} F_s(\theta)$, which is the comparison to the optimal constant prediction. This allows it to be robust to adversarial functions and adapted to non-stationary environments where very few assumptions can be made. Note here that the regret can be negative. This is presented in Section 11.1.

- **Partial feedback (zero-th order):** Independently of the regret framework, the feedback given to the algorithm may be less precise than the full gradient (e.g., only the function value). This is crucial in applications where function values are expensive to obtain without access to gradients.

  This is the domain of zero-th order optimization, which can be treated through gradient-based algorithms (Section 11.2) or the framework of multi-armed bandits (Section 11.3).

In this chapter, we briefly cover three topics from this large body of literature. For more details, see Shalev-Shwartz (2011); Bubeck and Cesa-Bianchi (2012); Hazan (2022); Slivkins (2019); Lattimore and Szepesvári (2020). This chapter aims to give the main ideas, how they differ from classical learning theory (using the unified notations we provide in this book), and to encourage readers to study these references.

11.1 First-order online convex optimization

In this section, we consider a sequence of arbitrary deterministic real-valued convex functions $F_s : \mathbb{R}^d \to \mathbb{R}$, $s \geqslant 1$, and a compact convex set $\mathcal{C}$. The goal of online convex optimization is, starting from a certain $\theta_0 \in \mathcal{C}$, to obtain a sequence $(\theta_s)_{s\geqslant 1}$ so that the

regret at time t, defined as

$$\frac{1}{t}\sum_{s=1}^{t} F_s(\theta_{s-1}) - \inf_{\theta\in\mathcal{C}} \frac{1}{t}\sum_{s=1}^{t} F_s(\theta), \tag{11.1}$$

is as small as possible.

We assume that at time s, we can access a gradient of F_s at any point $\theta_{s-1} \in \mathcal{C}$ that depends on past information. We also consider the possibility that we only observe a random, unbiased version g_s, that is, if $\mathcal{F}_s$ denotes the information up to (and including) time s,

$$\mathbb{E}\big[g_s|\mathcal{F}_{s-1}\big] = F_s'(\theta_{s-1}).$$

Given the added randomness, we consider the expected regret as a criterion.

For simplicity, we assume that almost surely, $\|g_s\|_2^2 \leqslant B^2$ (which in the context of machine learning corresponds to Lipschitz-continuous loss functions, which include the logistic loss, the hinge loss, and the square loss since we have assumed that we optimize on a bounded set[1]).

Applications. This is adapted to a non-stationary environment, where the data distribution varies over time, either stochastically or even adversarially (based on earlier predictions).

In this section, we only present the non-smooth case. The smooth case will be proposed as exercises but leads to similar results compared to the regular stochastic case.

11.1.1 Convex case

We consider the projected stochastic gradient descent recursion:

$$\theta_s = \Pi_{\mathcal{C}}(\theta_{s-1} - \gamma_s g_s),$$

for a certain positive step-size γ_s (which we assume deterministic for simplicity), where $\Pi_{\mathcal{C}}$ is the orthogonal projection onto the set $\mathcal{C}$. We then have, for any $\theta \in \mathcal{C}$ (as opposed to a fixed $\theta = \eta_*$ the global optimum, like in regular optimization in Chapter 5),

$$\begin{aligned}
\|\theta_s - \theta\|_2^2 &\leqslant \|\theta_{s-1}-\theta\|_2^2 - 2\gamma_s g_s^\top(\theta_{s-1}-\theta) + \gamma_s^2 B^2 \text{ by contractivity of projections,}\\
\mathbb{E}\big[\|\theta_s-\theta\|_2^2\big|\mathcal{F}_{s-1}\big) &\leqslant \|\theta_{s-1}-\theta\|_2^2 - 2\gamma_s F_s'(\theta_{s-1})^\top(\theta_{s-1}-\theta) + \gamma_s^2 B^2,\\
&\qquad\qquad \text{using the unbiasedness of the gradient,}\\
&\leqslant \|\theta_{s-1}-\theta\|_2^2 - 2\gamma_s\big[F_s(\theta_{s-1}) - F_s(\theta)\big] + \gamma_s^2 B^2, \text{ using convexity.}
\end{aligned}$$

Taking full expectations and isolating $F_s(\theta_{s-1}) - F_s(\theta)$, we get:

$$\mathbb{E}\big[F_s(\theta_{s-1}) - F_s(\theta)\big] \leqslant \frac{1}{2\gamma_s}\Big(\mathbb{E}\big[\|\theta_{s-1}-\theta\|_2^2\big] - \mathbb{E}\big[\|\theta_s-\theta\|_2^2\big]\Big) + \frac{\gamma_s}{2}B^2.$$

[1]The square loss is not Lipschitz-continuous on an unbounded domain, but is, once constrained to a bounded domain.

We can then sum between $s = 1$ to $s = t$, to obtain

$$\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\big[F_s(\theta_{s-1})\big]-\frac{1}{t}\sum_{s=1}^{t}F_s(\theta)\leqslant\frac{1}{t}\sum_{s=1}^{t}\frac{1}{2\gamma_s}\Big(\mathbb{E}\big[\|\theta_{s-1}-\theta\|_2^2\big]-\mathbb{E}\big[\|\theta_s-\theta\|_2^2\big]\Big)+\frac{1}{t}\sum_{s=1}^{t}\frac{\gamma_s}{2}B^2.$$

At this point, the proof is exactly the same as the one of Prop. 5.7, with only the appearances of functions F_s that depend on s.

In Chapter 5 (that is, the proof of Prop. 5.7), we considered non-uniform averaging, which is not adapted to the online setting (because the regret is based on a uniform average). We could also use a constant step-size that depends on the horizon t (which then needs to be known in advance). By using Abel's summation formula (discrete integration by part), we can use a time-dependent step-size sequence (γ_s), as, using the notation $\delta_s = \mathbb{E}\big[\|\theta_s - \theta\|_2^2\big]$, and for decreasing step-sizes:

$$\begin{aligned}
\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\big[F_s(\theta_{s-1})\big]-\frac{1}{t}\sum_{s=1}^{t}F_s(\theta) &\leqslant \frac{1}{t}\sum_{s=1}^{t}\frac{1}{2\gamma_s}(\delta_{s-1}-\delta_s)+\frac{1}{t}\sum_{s=1}^{t}\frac{\gamma_s}{2}B^2 \\
&\qquad\qquad \text{from the last equation,} \\
&= \frac{1}{t}\sum_{s=1}^{t-1}\delta_s\big(\frac{1}{2\gamma_{s+1}}-\frac{1}{2\gamma_s}\big)+\frac{\delta_0}{2t\gamma_1}-\frac{\delta_t}{2t\gamma_t}+\frac{1}{t}\sum_{s=1}^{t}\frac{\gamma_s}{2}B^2 \\
&\qquad\qquad \text{using Abel's summation formula,} \\
&\leqslant \frac{1}{t}\sum_{s=1}^{t-1}\mathrm{diam}(\mathcal{C})^2\big(\frac{1}{2\gamma_{s+1}}-\frac{1}{2\gamma_s}\big)+\frac{\mathrm{diam}(\mathcal{C})^2}{2t\gamma_1}+\frac{1}{t}\sum_{s=1}^{t}\frac{\gamma_s}{2}B^2 \\
&\qquad\qquad \text{using that } \delta_s \leqslant \mathrm{diam}(\mathcal{C})^2 \text{ for all } s, \\
&= \frac{\mathrm{diam}(\mathcal{C})^2}{2t\gamma_t}+\frac{1}{t}\sum_{s=1}^{t}\frac{\gamma_s}{2}B^2.
\end{aligned}$$

By choosing $\gamma_s = \frac{\mathrm{diam}(\mathcal{C})}{B\sqrt{s}}$, we get, using the same inequalities as for the proof of Prop. 5.7:

$$\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}\big[F_s(\theta_{s-1})\big]-\frac{1}{t}\sum_{s=1}^{t}F_s(\theta)\leqslant\frac{3B\mathrm{diam}(\mathcal{C})}{2\sqrt{t}}. \tag{11.2}$$

This is exactly the expected regret and essentially the same bound as stochastic optimization in Section 5.4. Note that from such a bound, if all F_s's are equal, we can do an "online-to-batch" conversion using Jensen's inequality and exactly get the bound for regular projected stochastic gradient descent (which is no surprise, as the proof is essentially the same).

We show in Section 11.1.4 that the rate in Eq. (11.2) is, up to constants, the best possible over all Lipschitz-continuous functions over a compact set.

Exercise 11.1 *(♦) In the unconstrained online optimization with smooth functions, that is, assuming that each F_t is L-smooth, and $\mathcal{C} = \mathbb{R}^d$, provide a regret bound for online gradient descent.*

11.1.2 Strongly-convex case (♦)

Assuming strong-convexity (e.g., by adding $\frac{\mu}{2}\|\theta\|_2^2$ to the objective function), we will get a rate proportional to $B^2\log(k)/(\mu k)$. Indeed, assuming that the functions F_s are all μ-strongly-convex on $\mathcal{C}$. We can indeed modify the proof above with the step-size $\gamma_s = 1/(\mu s)$, to get (with modifications in red):

$$\begin{aligned}
\|\theta_s - \theta\|_2^2 &\leqslant \|\theta_{s-1} - \theta\|_2^2 - 2\gamma_s g_s^\top(\theta_{s-1} - \theta) + \gamma_s^2 B^2 \\
\mathbb{E}\big[\|\theta_s - \theta\|_2^2 \big| \mathcal{F}_{s-1}\big) &\leqslant \|\theta_{s-1} - \theta\|_2^2 - 2\gamma_s F_s'(\theta_{s-1})^\top(\theta_{s-1} - \theta) + \gamma_s^2 B^2 \\
&\leqslant \|\theta_{s-1} - \theta\|_2^2 - 2\gamma_s\big[F_s(\theta_{s-1}) - F_s(\theta) + \frac{\mu}{2}\|\theta_{s-1} - \theta\|_2^2\big] + \gamma_s^2 B^2.
\end{aligned}$$

Taking full expectations and isolating function values, we get:

$$\mathbb{E}\big[F_s(\theta_{s-1}) - F_s(\theta)\big] \leqslant \big(\frac{1}{2\gamma_s} - \frac{\mu}{2}\big)E\big[\|\theta_{s-1} - \theta\|_2^2\big] - \frac{1}{2\gamma_s}\mathbb{E}\big[\|\theta_s - \theta\|_2^2\big]\Big) + \frac{\gamma_s}{2}B^2.$$

We can then use the specific form of step-size to get

$$\mathbb{E}\big[F_s(\theta_{s-1}) - F_s(\theta)\big] \leqslant \frac{\mu}{2}(s-1)E\big[\|\theta_{s-1} - \theta\|_2^2\big] - \frac{\mu}{2}s\mathbb{E}\big[\|\theta_s - \theta\|_2^2\big] + \frac{1}{2\mu s}B^2.$$

Then, summing between $s = 1$ to $s = t$, we obtain, with a telescoping sum:

$$\frac{1}{t}\sum_{s=1}^t \mathbb{E}\big[F_s(\theta_{s-1})\big] - \frac{1}{t}\sum_{s=1}^t F_s(\theta) \leqslant \frac{1}{t}\sum_{s=1}^t \frac{1}{2\mu s}B^2 \leqslant \frac{1}{2\mu t}(1 + \log t),$$

using the classical $\log(t)$ upper bound on the harmonic series. Note the appearance of $\log(t)$, which would not be the case if we had used the step-size $\gamma_s = \frac{2}{s+1}$ like in Exercise 5.29 (but which would require a different averaging scheme with weights proportional to s). For online learning, it turns out that the logarithmic term is unavoidable (Hazan and Kale, 2014).

11.1.3 Online mirror descent (♦)

In this section, we extend the online stochastic gradient descent recursion analysis from Section 11.1.1 to the *online mirror descent* framework, which will apply to the regular stochastic case as well.

Mirror map. We assume given a "mirror map" $\Phi : \mathcal{C}_\Phi \to \mathbb{R}$, which is differentiable and μ-strongly convex (on the set $\mathcal{C} \subset \mathcal{C}_\Phi$) with respect to a norm $\|\cdot\|$, that is,

$$\Phi(\eta) \geqslant \Phi(\theta) + \Phi'(\theta)^\top(\eta - \theta) + \frac{\mu}{2}\|\eta - \theta\|^2, \quad \forall \eta, \theta \in \mathcal{C}.$$

We also assume that the gradient Φ' is a bijection from $\mathcal{C}_\Phi$ to $\mathbb{R}^d$. Classical examples are:

- Squared Euclidean norm: $\Phi(\theta) = \frac{1}{2}\|\theta\|_2^2$ with full domain, and norm $\|\cdot\| = \|\cdot\|_2$, with $\mu = 1$.

- Entropy: $\Phi(\theta) = \sum_{i=1}^{d} \theta_i \log \theta_i$ with domain $\mathcal{C}_\Phi = (\mathbb{R}_+^*)^d$, and norm $\|\cdot\| = \|\cdot\|_1$, with $\mu = 1$ (a result which is equivalent to Pinsker inequality[2]).
- Squared ℓ_p-norms: $\Phi(\theta) = \frac{1}{2}\|\theta\|_p^2$ with full domain, for $p \in (1, 2]$, and norm $\|\cdot\| = \|\cdot\|_p$, with $\mu = p - 1$ (see proof of strong-convexity by Ball et al., 2002).

See also Exercise 13.2 in Chapter 13 for an example of mirror map for matrices.

Online mirror descent. We consider the same set-up as the beginning of Section 11.1.1, that is, we have convex Lipschitz-continuous functions F_s, for $s \geqslant 1$, and we access an unbiased (sub)gradient g_s, that is, if $\mathcal{F}_s$ denotes the information up to (and including) time s,

$$\mathbb{E}\big[g_s | \mathcal{F}_{s-1}\big] = F_s'(\theta_{s-1}).$$

The online mirror descent iteration is defined by:

$$\theta_t = \arg\min_{\theta \in \mathcal{C}} \; g_t^\top (\theta - \theta_{t-1}) + \frac{1}{\gamma} D_\Phi(\theta, \theta_{t-1}), \tag{11.3}$$

where $\mathcal{C}$ is a compact convex set, $D_\Phi(\theta, \eta) = \Phi(\theta) - \Phi(\eta) - \Phi'(\eta)^\top(\theta - \eta)$ is the Bregman divergence associated with the mirror map Φ, and γ is a step-size. If $\mathcal{C} \subset \mathcal{C}_\Phi$, then the update is simply defined by $\Phi'(\theta_t) = \Phi(\theta_{t-1}) - \gamma g_t$.

Proposition 11.1 *Given the mirror descent recursion in Eq. (11.3), assume that each stochastic gradient has bounded expected squared norm $\mathbb{E}[\|g_s\|_*^2 | \mathcal{F}_{s-1}] \leqslant B$, for all $s \geqslant 1$. Then, for every $\theta \in \mathcal{C}$, we have:*

$$\frac{1}{t} \sum_{s=1}^{t} \mathbb{E}\big[F_s(\theta_{s-1}) - F_s(\theta)\big] \leqslant \frac{1}{\gamma t} D_\Phi(\theta, \theta_0) + \frac{B^2 \gamma}{2\mu}.$$

Proof The proof follows the same structure as for online stochastic gradient descent in Section 11.1.1. From the optimality conditions of the update in Eq. (11.3), we have $(\theta - \theta_t)^\top \big(\gamma g_t + \Phi'(\theta_t) - \Phi'(\theta_{t-1})\big) \geqslant 0$ for all $\theta \in \mathcal{C}$. Given $\theta \in \mathcal{C}$, we have:

$$\begin{aligned}
& D_\Phi(\theta, \theta_t) - D_\Phi(\theta, \theta_{t-1}) \\
= \; & \Phi(\theta_{t-1}) + \Phi'(\theta_{t-1})^\top(\theta - \theta_{t-1}) - \Phi(\theta_t) - \Phi'(\theta_t)^\top(\theta - \theta_t) \\
= \; & \Phi(\theta_{t-1}) - \Phi(\theta_t) + \Phi'(\theta_{t-1})^\top(\theta_t - \theta_{t-1}) + (\Phi'(\theta_{t-1}) - \Phi'(\theta_t))^\top(\theta - \theta_t) \\
\leqslant \; & \Phi(\theta_{t-1}) - \Phi(\theta_t) + \Phi'(\theta_{t-1})^\top(\theta_t - \theta_{t-1}) + \gamma g_t^\top(\theta - \theta_t) \\
= \; & -D_\Phi(\theta_t, \theta_{t-1}) - \gamma g_t^\top(\theta_{t-1} - \theta) - \gamma g_t^\top(\theta_t - \theta_{t-1}) \\
\leqslant \; & -\frac{\mu}{2}\|\theta_t - \theta_{t-1}\|^2 - \gamma g_t^\top(\theta_{t-1} - \theta) + \gamma \|g_t\|_* \cdot \|\theta_t - \theta_{t-1}\| \leqslant \frac{\|g_t\|_*^2 \gamma^2}{2\mu} - \gamma g_t^\top(\theta_{t-1} - \theta),
\end{aligned}$$

using the strong-convexity of Φ and the bound on gradients. By taking conditional expectation, we get:

$$\mathbb{E}\big[D_\Phi(\theta, \theta_t) - D_\Phi(\theta, \theta_{t-1}) \big| \mathcal{F}_{t-1}\big] \leqslant \frac{B^2 \gamma^2}{2\mu} - \gamma F_t'(\theta_{t-1})^\top(\theta_{t-1} - \theta). \tag{11.4}$$

[2] see https://en.wikipedia.org/wiki/Pinsker%27s_inequality.

This leads to the desired result by using a telescoping sum and the convexity property $F_t(\theta_{t-1}) - F_t(\theta) \leqslant F_t'(\theta_{t-1})^\top(\theta_{t-1} - \theta)$. ■

We can make the following observations:

- We can optimize for the step-size when the optimization horizon t is known. Indeed, for $D^2 = 2\sup_{\theta,\theta' \in \mathcal{C}} D_\Phi(\theta, \theta')$, and the choice $\gamma = D\sqrt{\mu}/(B\sqrt{t})$, this leads to the regret bound $DB/\sqrt{\mu t}$. Alternatively, decaying step-sizes can be used like for regular SGD (which corresponds precisely to the feature map $\Phi = \frac{1}{2}\|\cdot\|_2^2$).
- With the online-to-batch conversion, we also get the same bound when all F_t's are equal for the averaged iterate.
- A classical application is for the simplex $\mathcal{C} = \{\theta \in \mathbb{R}^d_+, \sum_{j=1}^d \theta_j = 1\}$, and the entropy feature map. The update becomes $\theta_t \propto \theta_{t-1} \circ \exp(-\gamma g_t)$ (where $\circ$ denotes the component-wise product), with then a normalization step to sum to one, which is a *multiplicative* update, and the regret bound is equal to $B\sqrt{2\log(d)}/\sqrt{t}$. This regret bound would be of order $\sqrt{d}$ (instead of $\sqrt{\log d}$) if the Euclidean feature map was used.

Exercise 11.2 (Stochastic mirror descent for ℓ_1-regularization) *In the context of Prop. 11.1, we consider equal functions $F_t = F$, and assume that $\mathbb{E}[\|g_s\|_\infty^2 | \mathcal{F}_{s-1}] \leqslant B^2$ for all $s \geqslant 1$, and that $\theta_0 = 0$. Show that using mirror descent with the mirror map $\Phi(\theta) = \frac{1}{2}\|\theta\|_p^2$ for $p \in (1, 2]$, we get, for the average iterate $\bar{\theta}_t = \frac{1}{t}\sum_{s=1}^{t-1}\theta_s$, the bound $\mathbb{E}\big[F(\bar{\theta}_t)\big] \leqslant F(\theta) + \frac{1}{2\gamma t}\|\theta\|_1^2 + \frac{B^2 d^{2-2/p}\gamma}{2(p-1)}$. For $d \geqslant 2$, show that with $p = 1 + \frac{1}{\log d}$, the last term is less than $2B^2\gamma \log d$, and, if θ_* is the minimizer of F, with an appropriate choice of γ, $\mathbb{E}\big[F(\bar{\theta}_t)\big] - F(\theta_*) \leqslant \frac{2B\|\theta_*\|_1\sqrt{\log d}}{\sqrt{t}}$.*

11.1.4 Lower bounds (♦♦)

In order to prove a lower bound, following Abernethy et al. (2008), we consider the set $\mathcal{C} = \{\theta \in \mathbb{R}^d, \ \|\theta\|_\infty \leqslant 1\}$, and the linear (hence convex) function $F_t^{(\varepsilon)} : \mathbb{R}^d \to \mathbb{R}$ defined as $F_t^{(\varepsilon)}(\theta) = \varepsilon_t^\top \theta$, for $\varepsilon_s \in \{-1, 1\}^d$ for all $s \in \{1, \dots, t\}$. The gradient vectors g_t are then simply equal to ε_t. We here have the exact deterministic gradient, with constants $B = \sqrt{d}$ and $\operatorname{diam}(\mathcal{C}) = 2\sqrt{d}$.

To obtain a lower bound of performance, it suffices to show that for any sequence (θ_s),

$$\sup_{\varepsilon \in \mathcal{E}} \ \frac{1}{t}\sum_{s=1}^t F_s^{(\varepsilon)}(\theta_{s-1}) - \inf_{\theta \in \mathcal{C}} \frac{1}{t}\sum_{s=1}^t F_s^{(\varepsilon)}(\theta)$$

is lower-bounded for $\mathcal{E}$ a well-chosen set. As already used in proving lower bounds in Section 3.7 and in Chapter 15, this is lower-bounded by the expectation for any distribution on $\mathcal{E}$, which we take to be all independent Rademacher random variables (note that the algorithm is deterministic, with no noise in the gradients, but the problem itself is random).

The regret of any algorithm is $\frac{1}{t}\sum_{s=1}^t \varepsilon_s^\top \theta_{s-1}$, which has zero expectation because θ_{s-1} does not use the information of ε_s. Moreover, using that the ℓ_1-norm is dual to the ℓ_∞-norm:

$$\inf_{\theta\in\mathcal{C}} \frac{1}{t}\sum_{s=1}^t \varepsilon_s^\top \theta \quad = \quad -\Big\|\frac{1}{t}\sum_{s=1}^t \varepsilon_s\Big\|_1 = -d\Big|\frac{1}{t}\sum_{s=1}^t (\varepsilon_s)_1\Big|.$$

Therefore, from the following lemma, the regret is greater than $d\big|\frac{1}{t}\sum_{s=1}^t(\varepsilon_s)_1\big| \geqslant d/(8\sqrt{t})$, which is equal to $B\mathrm{diam}(\mathcal{C})/(16\sqrt{t})$, a lower bound that matches the upper-bound from SGD from Eq. (11.2), up to a constant factor.

Lemma 11.1 (Khintchine's inequality) *Let $\eta \in \{-1,1\}^t$ be a vector of independent Rademacher random variables (with equal probabilities for -1 and $+1$) and $x \in \mathbb{R}^d$. Let $p \in [0,\infty)$. Then*

$$\big(\mathbb{E}\big[|x^\top \eta|^p\big]\big)^{1/p} \leqslant B_p \|x\|_2, \tag{11.5}$$

with $B_p = \big(p 2^{p/2}\Gamma(p/2)\big)^{1/2}$, where Γ is the Gamma function.[3] *The bound B_p is less than $3\sqrt{p}$ for $p \geqslant 1$ and $\frac{3}{2}\sqrt{p}$ for $p \geqslant 2$. Moreover, if $p \geqslant 2$, $\big(\mathbb{E}\big[|x^\top\eta|^p\big]\big)^{1/p} \geqslant \|x\|_2$, and if $p \leqslant 2$, we have:*

$$\big(\mathbb{E}\big[|x^\top \eta|^p\big]\big)^{1/p} \geqslant B_{4-p}^{2-p/2}\|x\|_2. \tag{11.6}$$

We also have when $p \geqslant 1$, with $1/p + 1/q = 1$, $\big(\mathbb{E}\big[|x^\top\eta|^p\big]\big)^{1/p} \geqslant B_q^{-1}\|x\|_2$.

Proof (♦) We have, for $s = x^\top\eta$, and $p > 0$:

$$\mathbb{E}[|s|^p] = p\int_0^{+\infty} \lambda^{p-1}\mathbb{P}(|s| \geqslant \lambda) d\lambda,$$

(which can be checked using Fubini's theorem). We then compute directly:

$$\mathbb{E}[e^{ts}] = \prod_{i=1}^d \Big(\frac{1}{2}e^{ts} + \frac{1}{2}e^{-ts}\Big) = \prod_{i=1}^d \cosh(tx_i) \leqslant \exp(t^2\|x\|_2^2/2),$$

using that $\cosh\alpha \leqslant \exp(\alpha^2/2)$ for any $\alpha \in \mathbb{R}$. Thus, for $\lambda \geqslant 0$,

$$\begin{aligned}\mathbb{P}(|s| \geqslant \lambda) &\leqslant 2\mathbb{P}(s \geqslant \lambda) = 2\mathbb{P}(e^{ts} \geqslant e^{t\lambda}) \leqslant 2\inf_{t\geqslant 0} e^{-\lambda t}\mathbb{E}[e^{ts}] \text{ using Markov's inequality,}\\ &\leqslant 2\inf_{t\geqslant 0} e^{-\lambda t}\exp(t^2\|x\|_2^2/2) = 2\exp(-\lambda^2/(2\|x\|_2^2)), \text{ with } t = \lambda/\|x\|_2.\end{aligned}$$

Thus, through the change of variable $\mu = \lambda/\|x\|_2$:

$$\mathbb{E}[|s|^p] \leqslant 2p\int_0^{+\infty} \lambda^{p-1}\exp(-\lambda^2/(2\|x\|_2^2))d\lambda = \|x\|_2^p \times 2p\int_0^{+\infty}\mu^{p-1}\exp(-\mu^2/2)d\mu.$$

[3] See https://en.wikipedia.org/wiki/Gamma_function.

Thus, for Eq. (11.5), we can take $B_p^p = 2p\int_0^{+\infty}\lambda^{p-1}e^{-\lambda^2/2}d\lambda = p2^{p/2}\int_0^{+\infty}u^{p/2-1}e^{-u}du = p2^{p/2}\Gamma(p/2)$, with the change of variable $u = \lambda^2/2$. Through Stirling formula $\Gamma(p/2)^{1/p} \sim \sqrt{p/(2e)}$, and thus $B_p \sim \sqrt{p/e}$, and one can then check the bound $B_p \leqslant 3\sqrt{p}$ for $p \geqslant 1$, and $B_p \leqslant \frac{3}{2}\sqrt{p}$ for $p \geqslant 2$.

Assuming $\|x\|_2 = 1$ without loss of generality, we have, using Hölder's inequality:

$$1 = \mathbb{E}[|x^\top\eta|^2] \leqslant \big(\mathbb{E}[|x^\top\eta|^p]\big)^{1/p}\big(\mathbb{E}[|x^\top\eta|^q]\big)^{1/q},$$

which leads to the last lower bound.

Moreover, for $p \geqslant 2$, we have directly $\|x\|_2 \leqslant \big(\mathbb{E}\big[|x^\top\eta|^p\big]\big)^{1/p}$, and to prove Eq. (11.6), for $p \in [0,2]$, we have by Cauchy-Schwarz inequality:

$$\begin{aligned} 1 &= \mathbb{E}[|x^\top\eta|^2] = \mathbb{E}[|x^\top\eta|^{p/2}|x^\top\eta|^{2-p/2}] \leqslant \big(\mathbb{E}[|x^\top\eta|^p]\big)^{1/2}\big(\mathbb{E}[|x^\top\eta|^{4-p}]\big)^{1/2} \\ &\leqslant \big(\mathbb{E}[|x^\top\eta|^p]\big)^{1/2}B_{4-p}^{2-p/2}. \end{aligned}$$

The optimal constant in Eq. (11.5) is $B_p = 1$ for $p \in (0,2]$, and $B_p = \sqrt{2}(\Gamma(p/2+1/2)/\sqrt{\pi})^{1/p}$ if $p > 2$, while the optimal constant in Eq. (11.6) is $A_p = 1$ if $p \geqslant 1$, and $A_p = 2^{1/2-1/p}$ if $p < 1.847$.[4] ■

Thus, with the notations of the lemma above:

$$\big(\mathbb{E}[|x^\top\eta|^p]\big)^{1/p} \geqslant B_{4-p}^{1-4/p}\|x\|_2.$$

This leads to $\mathbb{E}|x^\top\eta| \geqslant \|x\|_2 B_3^{-3} \geqslant \|x\|_2(3\cdot 2^{3/2}\Gamma(3/2))^{-1} \geqslant \|x\|_2/8$ for the lower bound for online learning.

Exercise 11.3 *(♦) What would upper and lower bounds be if the regret criterion is replaced by* $\mathbb{E}\big[\sum_{s=1}^t \alpha_s F_s(\theta_{s-1})\big] - \inf_{\theta\in\mathcal{C}}\frac{1}{t}\sum_{s=1}^t \alpha_s F_s(\theta)$ *for an arbitrary sequence* (α_s) *of positive numbers?*

11.2 Zero-th order convex optimization

In this section, we consider the task of unconstrained minimization of a convex function F, given only access to function values, which is typically referred to as *zero-th order optimization* (since the function value is the zero-th order derivative of F, while the gradient is the vector of first-order derivatives).

If the function values are accessible with no noise and the function is smooth, then one can get a gradient by finite differences by defining the following estimate:

$$\hat{F}'(\theta) = \sum_{i=1}^d \frac{1}{\delta}\big[F(\theta+\delta e_i) - F(\theta)\big]e_i \in \mathbb{R}^d, \tag{11.7}$$

[4] See https://en.wikipedia.org/wiki/Khintchine_inequality.

where $(e_i)_{i\in\{1,\dots,d\}}$ is the canonical orthonormal basis of $\mathbb{R}^d$, with arbitrary precision when δ tends to zero. Indeed, using the smoothness inequality from Eq. (5.10):

$$\|\hat{F}'(\theta) - F'(\theta)\|_2^2 = \frac{1}{\delta^2}\sum_{i=1}^{d}\big[F(\theta+\delta e_i) - F(\theta) - F'(\theta)^\top \delta e_i\big]^2 \leqslant \frac{d}{\delta^2}\big(L\delta^2/2\big)^2 = \frac{dL^2\delta^2}{4}.$$

Therefore, assuming for simplicity that algorithms have infinite numerical precision at the expense of $d+1$ noiseless function evaluations (one at θ, and d at each $\theta+\delta e_i$), we can compute the exact gradient, and use gradient descent. Note also that for many functions, the gradient can be computed easily with automatic differentiation techniques (see, e.g., Baydin et al., 2018, and references therein). The problem is more interesting with noisy evaluations.

In this section, we first consider for simplicity the case where f is convex and smooth (that is, essentially with bounded second-order derivatives) but only accessible with a stochastic first-order oracle (unbiased, with variance σ^2), for which, in Eq. (11.7), the noise in the function values explodes when δ goes to zero.

That is, we will consider the iteration

$$\theta_t = \theta_{t-1} - \gamma\big[\frac{1}{\delta}\big(F(\theta_{t-1}+\delta z_t) + \zeta_t - F(\theta_{t-1}) - \zeta_t'\big)z_t\big],$$

where ζ_t and ζ_t' are zero-mean random variables with variance σ^2, corresponding to the additive noise on the two function evaluations. By writing $\varepsilon_t = \zeta_t - \zeta_t'$, we get:

$$\theta_t = \theta_{t-1} - \gamma\big[\frac{1}{\delta}\big(F(\theta_{t-1}+\delta z_t) - F(\theta_{t-1}) + \varepsilon_t\big)z_t\big], \tag{11.8}$$

where ε_t corresponds to the noise with the two function evaluations at θ_{t-1} and $\theta_{t-1}+\delta z_t$, thus of variance $2\sigma^2$, and z_t is sampled from a distribution so that the mean is $\mathbb{E}[z_t] = 0$ and the covariance matrix is $\mathbb{E}[z_t z_t^\top] = I$.

There are two natural candidates: (1) z a signed canonical basis vectors selected uniformly at random (that is, $\pm\sqrt{d}e_i$, with i selected uniformly at random in $\{1,\dots,d\}$, and a factor $\sqrt{d}$ to obtain an identity covariance matrix), which corresponds to a single coordinate change like in Eq. (11.7), or (2) z standard Gaussian vector (with mean zero and identity covariance matrix). We consider the second option, as this will lead to an interesting property relating the stochastic gradient estimate to the gradient of a modified function.

Note that if F is defined as an expectation $F(\theta) = \mathbb{E}_\xi\big[f(\theta,\xi)\big]$, the stochasticity at time t comes from a sample ξ_t. We then compute the function values $f(\theta,\xi_t)$ at *two* different points with the same ξ_t, and we can get an improved bound (see the end of Section 11.2.1).

The key in analyzing the iteration in Eq. (11.8) is to study $g = \frac{1}{\delta}\big(F(\theta+\delta z) - F(\theta)\big)z$, for a certain θ and z and a standard Gaussian vector.

For δ small, a simple Taylor expansion around θ leads to

$$g = \frac{1}{\delta}\big(F(\theta+\delta z) - F(\theta)\big)z = \frac{1}{\delta}\big(\delta z^\top F'(\theta) + O(\delta^2)\big)z = zz^\top F'(\theta) + O(\delta).$$

Thus, by taking an expectation with respect to z, we get $\mathbb{E}[g] = F'(\theta) + O(\delta)$, that is, we have an almost unbiased gradient (for δ small), and we can thus expect to use stochastic gradient techniques. It turns out that the analysis will be made even simpler through integration by parts and the property of the Gaussian distribution.

In terms of variance linked to noisy evaluations, the term $\frac{1}{\delta}\varepsilon_t z_t$ has zero mean, but its squared norm has expectation $\mathbb{E}\big[\big\|\frac{1}{\delta}\varepsilon_t z_t\big\|_2^2\big] = \frac{1}{\delta^2}2\sigma^2 d$. Thus, it explodes when δ goes to zero, thus leading to some trade-offs we now look at.

11.2.1 Smooth stochastic gradient descent

For simplicity, we consider an L-smooth function F defined on $\mathbb{R}^d$ (see next section for the non-smooth version).

An important tool will be to consider the function $F_\delta : \mathbb{R}^d \to \mathbb{R}$ defined as

$$F_\delta(\theta) = \mathbb{E}_{z\sim\mathcal{N}(0,I)}[F(\theta+\delta z)], \tag{11.9}$$

which is the expectation of F taken at point distributed as a Gaussian with mean θ and covariance matrix $\delta^2 I$.

Approximation properties. We can analyze the difference between F and F_δ when F is L-smooth:

$$\forall \theta \in \mathbb{R}^d,\ F_\delta(\theta) - F(\theta) = \mathbb{E}_{z\sim\mathcal{N}(0,I)}\big[F(\theta+\delta z) - F(\theta) - \delta F'(\theta)^\top z\big].$$

Thus, using Jensen's inequality, we get $F_\delta(\theta) \geqslant F(\theta)$ and using the smoothness bound from Eq. (5.10), we get:

$$\forall \theta \in \mathbb{R}^d,\ 0 \leqslant F_\delta(\theta) - F(\theta) \leqslant \frac{L\delta^2}{2}\mathbb{E}_{z\sim\mathcal{N}(0,I)}[\|z\|_2^2] = \frac{L}{2}\delta^2 d. \tag{11.10}$$

Moreover, we can compute the expectation of the squared norm of the gradient estimate

$$\begin{aligned}
& \mathbb{E}\Big[\big\|\frac{1}{\delta}\big(F(\theta+\delta z) - F(\theta)\big)z\big\|_2^2\Big] \\
\leqslant\ & 2\mathbb{E}\Big[\big\|\frac{1}{\delta}\big(F(\theta+\delta z) - F(\theta) - \delta F'(\theta)^\top z\big)z\big\|_2^2\Big] + 2\mathbb{E}\big[\|zz^\top F'(\theta)\|_2^2\big] \\
\leqslant\ & 2\mathbb{E}\big[\frac{L^2\delta^2}{4}\|z\|_2^6\big] + 2F'(\theta)^\top\mathbb{E}\big[\|z\|_2^2 zz^\top\big]F'(\theta) \text{ using smoothness,} \\
=\ & \frac{L^2\delta^2}{2}d(d+2)(d+4) + 2\|F'(\theta)\|_2^2 \cdot 3d \leqslant \frac{L^2\delta^2}{2}15d^3 + 6d\|F'(\theta)\|_2^2,
\end{aligned} \tag{11.11}$$

where we have used that $\|z\|_2^2$ is a chi-squared random variable and that we get in closed form $\mathbb{E}[\|z\|_2^6] = d(d+2)(d+4)$ and $\mathbb{E}\big[\|z\|_2^2 zz^\top\big] = 3dI$ (see the exercise below).

Exercise 11.4 *Show that for a standard Gaussian vector $z \in \mathbb{R}^d$ (with zero mean and covariance matrix identity), then $\mathbb{E}[\|z\|_2^6] = d(d+2)(d+4)$ and $\mathbb{E}\big[\|z\|_2^2 z z^\top\big] = 3dI$.*

Stochastic gradient descent. We can now analyze gradient descent and take conditional expectations given the information $\mathcal{F}_{s-1}$ up to time $s-1$, and use the standard manipulations from Chapter 5, starting from:

$$\theta_s - \theta_* = \theta_{s-1} - \theta_* - \gamma \frac{1}{\delta}\big(F(\theta_{s-1} + \delta z_s) - F(\theta_{s-1})\big) z_s - \frac{\gamma}{\delta}\varepsilon_s z_s,$$

to get, by expanding the squared norm:

$$\begin{aligned}
&\mathbb{E}\big[\|\theta_s - \theta_*\|_2^2 | \mathcal{F}_{s-1}\big] \\
\leqslant\ & \|\theta_{s-1} - \theta_*\|_2^2 - 2\gamma F_\delta'(\theta_{s-1})^\top(\theta_{s-1} - \theta_*) \\
&\qquad + 2\gamma^2 \mathbb{E}\big[\big\|\frac{1}{\delta}\big(F(\theta_{s-1} + \delta z_s) - F(\theta_{s-1})\big) z_s\big\|_2^2 \big| \mathcal{F}_{s-1}\big] + 2\frac{\gamma^2}{\delta^2}\mathbb{E}[\varepsilon_s^2 \|z_s\|_2^2] \\
\leqslant\ & \|\theta_{s-1} - \theta_*\|_2^2 - 2\gamma F_\delta'(\theta_{s-1})^\top(\theta_{s-1} - \theta_*) \\
&\qquad + 2\gamma^2 \cdot \big[\frac{L^2\delta^2}{2} 15 d^3 + 6d \|F'(\theta_{s-1})\|_2^2\big] + 2\frac{\gamma^2}{\delta^2} \cdot 2d\sigma^2 \text{ using Eq. (11.11)}, \\
\leqslant\ & \|\theta_{s-1} - \theta_*\|_2^2 - 2\gamma\big[F_\delta(\theta_{s-1}) - F_\delta(\theta_*)\big] \\
&\qquad + 15\gamma^2 L^2 \delta^2 d^3 + 24 L\gamma^2 d\big[F(\theta_{s-1}) - F(\theta_*)\big] + 4d\frac{\gamma^2}{\delta^2}\sigma^2 \text{ using co-coercivity}, \\
\leqslant\ & \|\theta_{s-1} - \theta_*\|_2^2 - 2\gamma\big[F(\theta_{s-1}) - F(\theta_*)\big] + 2\gamma \cdot \frac{L}{2}\delta^2 d \\
&\qquad + 15\gamma^2 L^2 \delta^2 d^3 + 24 L\gamma^2 d\big[F(\theta_{s-1}) - F(\theta_*)\big] + 4d\frac{\gamma^2}{\delta^2}\sigma^2, \text{ using Eq. (11.10)}.
\end{aligned}$$

Thus, if $\gamma \leqslant \frac{1}{24dL}$, we have $24L\gamma^2 d \leqslant \gamma$, and we get:

$$\begin{aligned}
\mathbb{E}\big[\|\theta_s - \theta_*\|_2^2 | \mathcal{F}_{s-1}\big] &\leqslant \|\theta_{s-1} - \theta_*\|_2^2 - \gamma\big[F(\theta_{s-1}) - F(\theta_*)\big] + \gamma L\delta^2 d \\
&\qquad + \frac{15}{24}\gamma L\delta^2 d^2 + 4d\frac{\gamma^2}{\delta^2}\sigma^2 \\
&\leqslant \|\theta_{s-1} - \theta_*\|_2^2 - \gamma\big[F(\theta_{s-1}) - F(\theta_*)\big] + 2\gamma L\delta^2 d^2 + 4d\frac{\gamma^2}{\delta^2}\sigma^2,
\end{aligned}$$

leading to, taking full expectations:

$$\mathbb{E}\big[F(\theta_{s-1})\big] - F(\theta_*) \leqslant \frac{1}{\gamma}\Big(\mathbb{E}\big[\|\theta_{s-1} - \theta_*\|_2^2\big] - \mathbb{E}\big[\|\theta_s - \theta_*\|_2^2\big]\Big) + 2L\delta^2 d^2 + 4d\frac{\gamma}{\delta^2}\sigma^2.$$

Summing from $s = 1$ to $s = t$, we get

$$\frac{1}{t}\sum_{s=1}^t \mathbb{E}\big[F(\theta_{s-1})\big] - F(\theta_*) \leqslant \frac{1}{\gamma t}\|\theta_0 - \theta_*\|_2^2 + 2L\delta^2 d^2 + 4d\frac{\gamma}{\delta^2}\sigma^2. \tag{11.12}$$

We can now analyze various situations depending on the presence or absence of noise (see empirical illustration in Figure 11.1):

- If $\sigma = 0$, then we can take δ as close to zero as possible and get the rate, with $\gamma = \frac{1}{24dL}$, for the average iterate $\bar{\theta}_t = \frac{1}{t}\sum_{s=1}^{t} \theta_{s-1}$, and using Jensen's inequality:

$$\mathbb{E}\big[F(\bar{\theta}_t)\big] - F(\theta_*) \leqslant \frac{24Ld}{t}\|\theta_0 - \theta_*\|_2^2. \tag{11.13}$$

  As suggested at the beginning of Section 11.2, we only lose a factor of d compared to regular gradient descent in Section 5.2.4.

- If $\sigma > 0$, we can optimize over δ to get (assuming σ is known), with the choice $\delta^4 = 2\gamma\sigma^2 L^{-1} d^{-1}$, $\mathbb{E}\big[F(\bar{\theta}_t)\big] - F(\theta_*) \leqslant \frac{1}{\gamma t}\|\theta_0 - \theta_*\|_2^2 + 2\sqrt{2} \cdot \gamma^{1/2} L^{1/2} \sigma d^{3/2}$. With the maximal allowed step-size $\gamma = \frac{1}{24dL}$, this leads to

$$\mathbb{E}\big[F(\bar{\theta}_t)\big] - F(\theta_*) \quad \leqslant \quad \frac{24Ld}{t}\|\theta_0 - \theta_*\|_2^2 + \sigma d.$$

  There is convergence only up to the noise level with a limiting bound σd. We can also use a step-size γ that depends on the horizon t, by taking $\gamma = \frac{1}{24Ld} t^{-2/3}$, leading to:

$$\mathbb{E}\big[F(\bar{\theta}_t)\big] - F(\theta_*) \quad \leqslant \quad \frac{d}{t^{1/3}}\big[24L\|\theta_0 - \theta_*\|_2^2 + \sigma\big].$$

  We not only lose a factor of d in the bound, but the dependence in t is worsened from $1/t$ to $1/t^{1/3}$. Note that (a) the natural rate for convex stochastic first-order methods is $O(1/\sqrt{t})$, and (b) the dependence in σ could be improved if the noise level were known.

Extensions. We can also consider the case where we can do two function evaluations, where one can check that we can essentially remove the variance term in $d\frac{\gamma^2}{\delta^2}\sigma^2$ due to two noisy evaluations, removing in Eq. (11.12) the last term, and thus with improved behavior. For related lower bounds, see Duchi et al. (2015).

Exercise 11.5 *When two function evaluations are available, compute optimal values of δ and γ and provide an improved convergence rate.*

11.2.2 Stochastic smoothing (♦)

In this section, we consider the case where F may not be smooth, which leads to considering the nice effect of randomized smoothing. This randomized smoothing can simply be explained by seeing F_δ as the convolution of the function F by the density of the Gaussian distribution with mean zero and covariance matrix $\delta^2 I$. Since this density is infinitely differentiable, a continuous function will be turned into an infinitely differentiable function. One particular instance of this phenomenon is shown precisely below.

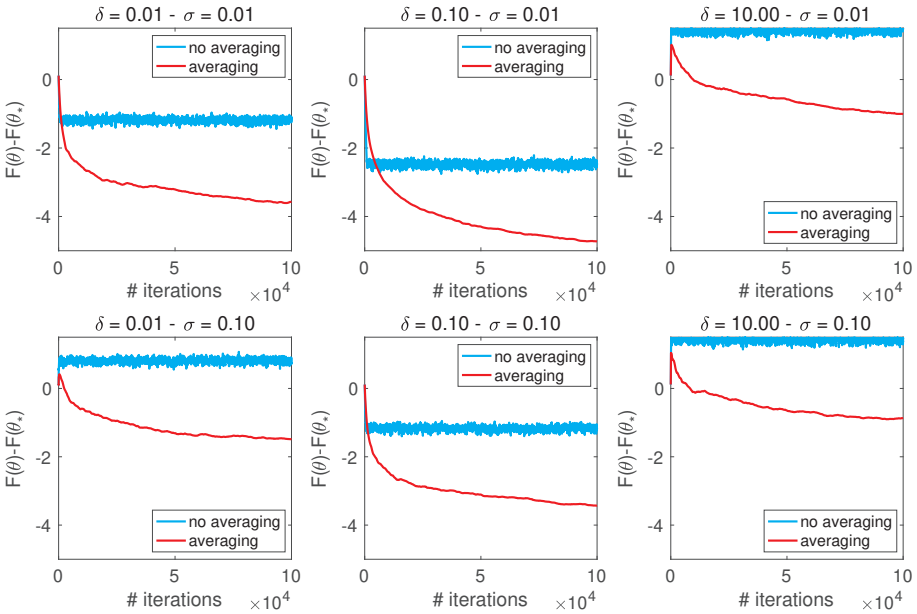
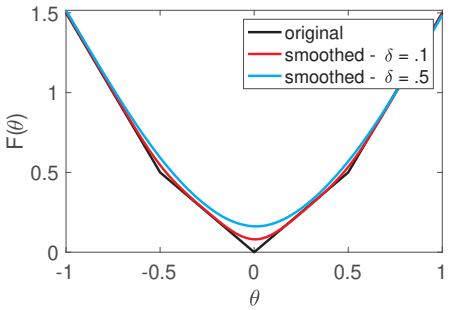


Figure 11.1: Zero-th order optimization with Gaussian smoothing on a quadratic function F in dimension $d = 10$, with step-size $\gamma = 1/(4Ld)$: two different levels of noise added to the function values, $\sigma = 0.01$ (top), and $\sigma = 0.1$ (bottom), with three different smoothing constants, $\delta = 0.01$ (left), $\delta = 0.1$ (middle), and $\delta = 10$ (right). Performance improves with smaller noise variance σ^2, while δ should be chosen not too large (then too much bias) and not too small (too much variance).



Proposition 11.2 (Randomized smoothing) *Assume F is B-Lipschitz-continuous. Then the function $F_\delta : \mathbb{R}^d \to \mathbb{R}$ defined in Eq. (11.9) is also B-Lipschitz-continuous. Morerover, it is $\big(\frac{\sqrt{d}}{\delta}B\big)$-smooth, with gradient equal to*

$$F'_\delta(\theta) = \frac{1}{\delta}\mathbb{E}_{z\sim\mathcal{N}(0,I)}\big[F(\theta+\delta z)z\big] = \frac{1}{\delta}\mathbb{E}_{z\sim\mathcal{N}(0,I)}\big[(F(\theta+\delta z)-F(\theta))z\big].$$

Moreover, $\forall\theta\in\mathbb{R}^d,\ |F_\delta(\theta)-F(\theta)| \leqslant B\delta\sqrt{d}$.

Proof If F is B-Lipschitz-continuous, then for any $\theta,\theta'\in\mathbb{R}^d$, we have

$$\begin{aligned}|F_\delta(\theta)-F_\delta(\theta')| &= \big|\mathbb{E}\big[F(\theta+\delta z)-F(\theta'+\delta z)\big]\big| \leqslant \mathbb{E}\big[|F(\theta+\delta z)-F(\theta'+\delta z)|\big]\\ &\leqslant \mathbb{E}\big[B\|\theta-\theta'\|_2\big] = B\|\theta-\theta'\|_2,\end{aligned}$$

which shows Lipschitz-continuity of F_δ. In terms of approximation, we have:

$$\forall\theta\in\mathbb{R}^d,\ |F_\delta(\theta)-F(\theta)| \leqslant \mathbb{E}_{z\sim\mathcal{N}(0,I)}\big[|F(\theta+\delta z)-F(\theta)|\big] \leqslant B\delta\mathbb{E}_{z\sim\mathcal{N}(0,I)}[\|z\|_2] \leqslant B\delta\sqrt{d}.$$

We can now use the expression of the multivariate standard Gaussian density to get:

$$F_\delta(\theta) = \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} F(\theta+\delta\eta)\exp\big(-\frac{1}{2}\|\eta\|_2^2\big)d\eta.$$

Then, assuming for simplicity that we can differentiate through the expectation, we get, by integration by parts:

$$\begin{aligned}F'_\delta(\theta) &= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} F'(\theta+\delta\eta)\exp\big(-\frac{1}{2}\|\eta\|_2^2\big)d\eta\\ &= \frac{1}{(2\pi)^{d/2}}\frac{1}{\delta}\int_{\mathbb{R}^d} \delta F'(\theta+\delta\eta)\exp\big(-\frac{1}{2}\|\eta\|_2^2\big)d\eta\\ &= \frac{1}{(2\pi)^{d/2}}\frac{1}{\delta}\int_{\mathbb{R}^d} \frac{\partial F(\theta+\delta\eta)}{\partial\eta}\exp\big(-\frac{1}{2}\|\eta\|_2^2\big)d\eta\\ &= -\frac{1}{(2\pi)^{d/2}}\frac{1}{\delta}\int_{\mathbb{R}^d} F(\theta+\delta\eta)\frac{\partial\exp\big(-\frac{1}{2}\|\eta\|_2^2\big)}{\partial\eta}d\eta \text{ by integration by parts,}\\ &= -\frac{1}{(2\pi)^{d/2}}\frac{1}{\delta}\int_{\mathbb{R}^d} F(\theta+\delta\eta)\exp\big(-\frac{1}{2}\|\eta\|_2^2\big)\big(-\eta\big)d\eta = \mathbb{E}\Big[\frac{1}{\delta}F(\theta+\delta z)z\Big].\end{aligned}$$

That is, the gradient is equal to

$$F'_\delta(\theta) = \mathbb{E}\Big[\frac{1}{\delta}F(\theta+\delta z)z\Big] = \mathbb{E}\Big[\frac{1}{\delta}\big(F(\theta+\delta z)-F(\theta)\big)z\Big].$$

The function F_δ is $\big(\frac{\sqrt{d}}{\delta}B\big)$-smooth, since for $\theta,\theta' \in \mathbb{R}^d$,

$$\|F'_\delta(\theta)-F'_\delta(\theta')\|_2 \leqslant \frac{1}{\delta}\mathbb{E}_{z\sim\mathcal{N}(0,I)}\big[|F(\theta+\delta z)-F(\theta'+\delta z)|\|z\|\big] \leqslant \frac{B}{\delta}\|\theta-\theta'\|_2\mathbb{E}_{z\sim\mathcal{N}(0,I)}[\|z\|_2].$$

■

In other words, the expectation of the gradient estimate happens to be exactly the gradient of a smoothed version F_δ of F. This will be used in the proof below. Moreover, the expression of F'_δ as an expectation leads naturally to the stochastic gradient $\hat{F}'_\delta(\theta) = \frac{1}{\delta}F(\theta+\delta z)z - \frac{1}{\delta}F(\theta)z$, for which we have: $\mathbb{E}[\hat{F}'_\delta(\theta)] = F'_\delta(\theta)$ and

$$\mathbb{E}\big[\|\hat{F}'_\delta(\theta)\|_2^2\big] \leqslant \mathbb{E}\big[B^2\|z\|_2^4\big] \leqslant 4B^2d^2.$$

Stochastic gradient descent. We have, for θ_* a minimizer of F on $\mathbb{R}^d$, by expanding the square,

$$\begin{aligned}\|\theta_s-\theta_*\|_2^2 &= \|\theta_{s-1}-\theta_*\|_2^2 - 2\frac{\gamma}{\delta}\big(\big[F(\theta_{s-1}+\delta z_s)-F(\theta_{s-1})+\varepsilon_s\big]z_s\big)^\top(\theta_{s-1}-\theta_*)\\ &\quad + \frac{\gamma^2}{\delta^2}\big\|\big[F(\theta_{s-1}+\delta z_s)-F(\theta_{s-1})+\varepsilon_s\big]z_s\big\|_2^2.\end{aligned}$$

We have, using the previous inequalities:

$$\mathbb{E}\big[\|\theta_s-\theta\|_2^2|\mathcal{F}_{s-1}\big] = \|\theta_{s-1}-\theta\|_2^2 - 2\gamma F'_\delta(\theta_{s-1})^\top(\theta_{s-1}-\theta) + 2\gamma^2\cdot 4B^2d^2 + 2\frac{\gamma^2}{\delta^2}\cdot\sigma^2 d,$$

leading to

$$\begin{aligned}F_\delta(\theta_{s-1})-F_\delta(\theta) &\leqslant \frac{1}{2\gamma}\Big(\mathbb{E}\big[\|\theta_{s-1}-\theta\|_2^2\big]-\mathbb{E}\big[\|\theta_s-\theta\|_2^2\big]\Big)+4\gamma B^2d^2+\frac{\gamma}{\delta^2}\sigma^2 d\\ F(\theta_{s-1})-F(\theta) &\leqslant \frac{1}{2\gamma}\Big(\mathbb{E}\big[\|\theta_{s-1}-\theta\|_2^2\big]-\mathbb{E}\big[\|\theta_s-\theta\|_2^2\big]\Big)+4\gamma B^2d^2+\frac{\gamma}{\delta^2}\sigma^2 d+2B\delta\sqrt{d}.\end{aligned}$$

We thus get

$$\frac{1}{t}\sum_{s=1}^t F(\theta_{s-1})-F(\theta) \leqslant \frac{1}{2\gamma t}\|\theta_0-\theta\|_2^2+4\gamma B^2d^2+\frac{\gamma}{\delta^2}\sigma^2 d+2B\delta\sqrt{d}.$$

This leads to a similar discussion as for the smooth case in Section 11.2.1, for the choice of step-sizes:

- When $\sigma = 0$ (no noise in function evaluations), we can take δ as small as possible so that rounding errors do not perturb the finite differences, and we then only lose a factor of d compared to the standard subgradient method studied in Section 5.3.

- When $\sigma > 0$, then we can optimize over δ, with $\delta^3 = \gamma\sigma^2 B^{-1}\sqrt{d}$. We then get

$$\frac{1}{t}\sum_{s=1}^{t} F(\theta_{s-1}) - F(\theta) \quad \leqslant \quad \frac{1}{2\gamma t}\|\theta_0 - \theta\|_2^2 + 4\gamma B^2 d^2 + 3d^{2/3}\gamma^{1/3}\sigma^{2/3}B^{2/3}.$$

To optimize the rate for large values of t, we can take $\gamma = \frac{1}{B^2 d^{1/2} t^{3/4}}$ for a final rate

$$\frac{d^{1/2}}{2t^{1/4}}\Big(B^2\|\theta_0 - \theta\|_2^2 + 6\sigma^{2/3}\Big) + 4\frac{d^{3/2}}{t^{3/4}}.$$

11.2.3 Extensions

In this section on zero-th order algorithms, we have focused on optimization algorithms with potentially stochastic noise, with a criterion which is the function values at the final time. This can be extended to online learning formulations with a different function F_t at time t, then using the regret criterion in Eq. (11.1). Online zero-th order optimization is significantly more complicated, and in the next section, we will focus only on multi-armed bandits, which are optimization problems over finite sets and already lead to significant theoretical and practical developments. For more general cases, see Hazan (2022).

11.3 Multi-armed bandits

This section aims to provide the simplest results for multi-armed stochastic bandits. There is extensive and rich literature; see Bubeck and Cesa-Bianchi (2012); Lattimore and Szepesvári (2020); Slivkins (2019) for a more detailed account.

Multi-armed bandits are the simplest model of sequential decision problems where information is gathered as decisions are made and losses incurred, where the "exploration-exploitation" dilemma occurs. Beyond being a stepping stone for many more complex models, it directly applies to clinical trials or routing in networks.

We consider k potential "arms" with associated means $\mu^{(1)}, \dots, \mu^{(k)} \in \mathbb{R}$. Every time we select the arm i, we receive a reward sampled independently of all other rewards and the previous arm choices from a sub-Gaussian distribution with mean $\mu^{(i)}$, and sub-Gaussian parameter σ. At time s, we select the arm i_s based on the information $\mathcal{F}_{s-1}$ up to time $s-1$ (that is, the rewards received before time $s-1$) and receive the reward r_s. In this chapter, we focus on "plain" bandits, noting that many variations exist where limited feedback is given to the algorithm, in particular "contextual" bandits, where a feature vector is observed before each arm is selected and where rewards are unknown functions of the feature vectors.

Criterion for reward maximization. Our criterion is the expected regret (adapted to the *maximization* of rewards), equal to

$$R_t = t \cdot \max_{i\in\{1,\dots,k\}} \mu^{(i)} - \sum_{s=1}^{t} \mathbb{E}[r_s].$$

As opposed to online learning in the previous section, here we are not dividing the regret by t.

Denoting $\Delta^{(j)} = \max_{i\in\{1,\dots,k\}} \mu^{(i)} - \mu^{(j)} \geqslant$ the difference between the mean of the best arm and the mean of arm j, and $n_t^{(j)}$ the number of times the arm j was selected in the first t iterations, we can express the regret as

$$R_t = \sum_{j=1}^{k} \Delta^{(j)} \mathbb{E}\big[n_t^{(j)}\big]. \tag{11.14}$$

Thus, the regret is a direct function of the number of times each arm is selected. For all algorithms, we consider the natural unbiased estimate of the arm means at time s, that is,

$$\hat{\mu}_t^{(j)} = \frac{1}{n_t^{(j)}} \sum_{s=1}^{t} r_s 1_{i_s=j} = \frac{1}{n_t^{(j)}} \sum_{a=1}^{n_t^{(j)}} x_a^{(j)},$$

where we imagine we select rewards from a sequence of i.i.d. samples $x_a^{(i)}$ with mean $\mu^{(i)}$ from each arm. This implies that as we select some arms multiple times, we get a more accurate estimate of $\mu^{(i)}$ as the expected squared distance between $\hat{\mu}_t^{(j)}$ and $\mu^{(i)}$ is proportional to $1/n_t^{(j)}$. To simplify the exposition, we ignore the equality cases among the various estimated $\hat{\mu}_t^{(j)}$, which is safe as long as the distributions of the arm values are absolutely continuous with respect to the Lebesgue measure.

11.3.1 Need for an exploration-exploitation trade-off

We can now consider two extreme algorithms, highlighting the need to both "explore" and "exploit".

Pure exploration. If we select a random arm at each step, then, from Eq. (11.14) and $\mathbb{E}[n_t^{(j)}] = \frac{t}{k}$, the expected regret is $t \cdot \frac{1}{k} \sum_{j=1}^{k} \Delta^{(j)}$ and depends linearly in t, that is, we have a "linear regret". At time step t, we get a reasonable estimate of the best arm, but this incurs a strong loss along the iterations.

Pure exploitation. The previous strategy was ignoring the online estimates $\hat{\mu}_t^{(j)}$. The pure exploitation strategy does the opposite by only selecting the arm with the current largest estimate, assuming that the first k steps are dedicated to selecting each arm only once. This has linear regret because there is a non-zero probability that the best arm will never be selected again.

Exercise 11.6 *Provide a lower bound on the regret of the pure exploitation strategy.*

11.3.2 "Explore-then-commit"

If we consider mk steps where we select exactly each arm m times, we can build the m estimates $\hat{\mu}^{(1)}, \dots, \hat{\mu}^{(k)}$, which are all independent random variables, with means $\mu^{(1)}, \dots, \mu^{(k)}$ and with sub-Gaussian parameters σ^2/m. Let i_* be the optimal arm.

We then select the arm with maximal $\hat{\mu}^{(j)}_{mk}$ for all remaining $t - km$ steps. The regret for this algorithm is then equal to, using Eq. (11.14), for $t > mk$:

$$R_t = m\sum_{j=1}^{k} \Delta^{(j)} + (t-mk)\sum_{j=1}^{k} \Delta^{(j)} \mathbb{P}(\hat{\mu}^{(j)}_{mk} \geqslant \hat{\mu}^{(i)}_{mk}, \forall i \neq j),$$

where the first term corresponds to the first m steps, for which this is the exact contribution of the regret, and the second term corresponds to the other $(t - mk)$ steps, where the arm j is selected if $\hat{\mu}^{(i)}_{mk}$ is maximized for $i = j$.

We can now upper-bound the second term by only imposing that an arm j is selected if $\hat{\mu}^{(j)}_{mk} \geqslant \hat{\mu}^{(i_*)}_{mk}$ (noting that $\Delta^{(i_*)} = 0$):

$$\begin{aligned} R_t &\leqslant m\sum_{j=1}^{k} \Delta^{(j)} + (t-mk)\sum_{j=1}^{k} \Delta^{(j)} \mathbb{P}(\hat{\mu}^{(j)}_{mk} \geqslant \hat{\mu}^{(i_*)}_{mk}) \\ &\leqslant m\sum_{j \neq i_*} \Delta^{(j)} + t\sum_{j \neq i_*} \Delta^{(j)} \exp\Big(-\frac{(\Delta^{(j)})^2 m}{4\sigma^2}\Big), \end{aligned}$$

by using sub-Gaussian tail bounds (see Section 1.2.1) on the difference of the m arm values between j and i_* (that is, $\hat{\mu}^{(i_*)}_{mk} - \hat{\mu}^{(j)}_{mk}$ is a sub-Gaussian random variable with mean $\Delta^{(j)}$ and sub-Gaussianity parameter $2\sigma^2/m$).

Two arms ($k = 2$). For $k = 2$ arms, the upper-bound is, with $\Delta = \Delta^{(i)}$ for $i \neq i_*$:

$$m\Delta + t\Delta \exp\Big(-\frac{\Delta^2 m}{4\sigma^2}\Big),$$

and we can minimize approximately with respect to m by taking the gradient with respect to m (assuming for a moment it is not restricted to be an integer), leading to $\Delta = t\frac{\Delta^3}{4\sigma^2}\exp\big(-\frac{\Delta^2 m}{4\sigma^2}\big)$, that is, we consider the candidate $m^* = \big\lfloor \frac{4\sigma^2}{\Delta^2} \log \frac{\Delta^2 t}{4\sigma^2} \big\rfloor$.

If $t > \frac{4\sigma^2}{\Delta^2}\exp\big(\frac{4\sigma^2}{\Delta^2}\big)$, then $m^* \geqslant 1$, while it is always less than $t/2$. We then have a regret less than (using $\log \alpha \leqslant \alpha - 1$):

$$\begin{aligned} \frac{4\sigma^2}{\Delta}\log\frac{\Delta^2 t}{4\sigma^2} + t\Delta\exp\Big[-\frac{\Delta^2}{4\sigma^2}\Big(\frac{4\sigma^2}{\Delta^2}\log\frac{\Delta^2 t}{4\sigma^2} - 1\Big)\Big] &= \frac{4\sigma^2}{\Delta}\Big[\exp(\Delta^2/(4\sigma^2)) + 2\log\frac{\Delta\sqrt{t}}{2\sigma}\Big] \\ &\leqslant \frac{4\sigma^2}{\Delta}\Big(\exp(\Delta^2/(4\sigma^2)) - 2 + 2\frac{\Delta\sqrt{t}}{2\sigma}\Big), \end{aligned}$$

which is less than a constant plus $4\sigma\sqrt{t}$. As shown below, this simple algorithm will achieve the lower bound (up to constant factors) for all possible algorithms. However, this requires knowing Δ and t in advance to select m^* appropriately.

More than two arms ($k \geqslant 2$). We consider the event $\mathcal{A} = \{\forall i \neq i_*, \hat{\mu}^{(i)} - \mu^{(i)} \leqslant \frac{r}{\sqrt{m}},\ \hat{\mu}^{(i_*)} - \mu^{(i_*)} \geqslant -\frac{r}{\sqrt{m}}\}$, where r is a constant to be determined later. This event is true if suboptimal arms are not too overestimated while the optimal arm is not too underestimated. If the event $\mathcal{A}$ is true, then the loss in rewards for the last $t - mk$ steps is less than $2\frac{r}{\sqrt{m}}$ (since only arms with means that are less than $2\frac{r}{\sqrt{m}}$ away from the optimal one can be selected), while it is less than $\delta = \max_{i \neq i_*} \Delta^i$ otherwise. Moreover, using sub-Gaussian tail bounds and the union bound, $\mathbb{P}(\mathcal{A}^c) \leqslant k \exp(-\frac{r^2}{2\sigma^2})$.

Thus, the regret is less than

$$R_t \leqslant mk\delta + 2\frac{rt}{\sqrt{m}} + \delta kt \exp\Big(-\frac{r^2}{2\sigma^2}\Big),$$

where the first term corresponds to the explore phase and the last two terms to the commit phase.

With $m^{3/2} \approx rt/(k\delta)$, we can minimize the first two terms and get

$$R_t \leqslant 3(rt)^{2/3}(k\delta)^{1/3} + \delta kt \exp\Big(-\frac{r^2}{2\sigma^2}\Big).$$

With $r = \sigma\sqrt{2\log(kt)}$, we then get $R_t \leqslant \delta + 3t^{2/3}k^{1/3}\delta^{1/3}\sigma^{2/3}\big(2\log(kt)\big)^{1/3}$, which grows as $t^{2/3}$ and does not achieve the lower bound (see a better algorithm in Section 11.3.3).

ε-greedy. We can mix exploration and exploitation with the "ε-greedy" strategy, which will update estimates $\hat{\mu}^{(i)}$ but spread the exploration phase over iterations by selecting with some positive probability a random arm. The final regret is similar to explore-and-commit (Auer et al., 2002a).

11.3.3 Optimism in the face of uncertainty (♦)

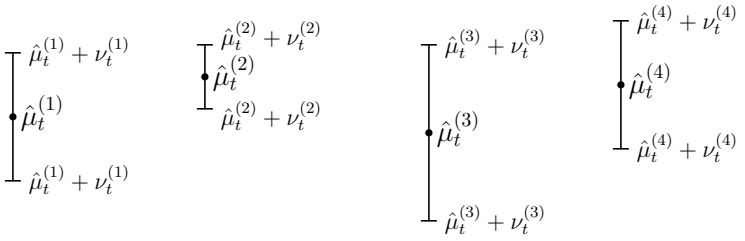
We consider the classical "upper confidence bound" (UCB) algorithm (Auer et al., 2002a), whose principle is simple. As arms are being selected, confidence intervals for the values of each arm are maintained as $[\hat{\mu}_t^{(i)} - \nu_t^{(i)}, \hat{\mu}^{(i)} + \nu_t^{(i)}]$. The arm that is selected is the one with maximal upper-confidence bound $\hat{\mu}_t^{(i)} + \nu_t^{(i)}$. This is one instance of the general principle of optimism in the face of uncertainty (Munos et al., 2014).

The precise algorithm is as follows (assuming that σ is known):

- For the first k rounds, select each arm exactly once, and form $\hat{\mu}_k^{(i)}$ as the reward received for arm i, with $\nu_k^{(i)} = \sqrt{2\rho\sigma^2\log(k)/n_k^{(i)}} = \sqrt{2\rho\sigma^2\log(k)}$, with $\rho > 0$ to be determined later.
- For all other $t > k$, select the arm i_t which maximizes $\hat{\mu}_{t-1}^{(i)} + \nu_{t-1}^{(i)}$, receive the reward, and update, for all i, $\hat{\mu}_t^{(i)}$ as the average reward received for all arms $i \in \{1, \dots, k\}$, with the interval width $\nu_t^{(i)} = \sqrt{2\rho\sigma^2\log(t)/n_t^{(i)}}$.

The confidence interval lenght for arm i is naturally proportional to $\sigma/\sqrt{n_t^{(i)}}$ with an extra factor that ensures sub-linear regret.

Thus, as illustrated below for $k = 4$, we have k confidence intervals, and we select the arm with the largest upper confidence bound (here $i = 4$).



The analysis consists in upper-bounding $\mathbb{E}\big[n_t^{(i)}\big]$ for $i \neq i_*$ and using Eq. (11.14), that is, $R_t = \sum_{i \neq i_*} \Delta^{(i)} \mathbb{E}\big[n_t^{(i)}\big]$, to obtain the regret bound. We follow the proof technique from Garivier and Cappé (2011). For simplicity, we assume that there is a single arm i_* with maximal mean.

The main idea of the proof is to compare the upper-confidence bounds to the optimal arm mean $\mu^{(i_*)}$. That is, for $i \neq i_*$, we have:

$$\begin{aligned}
\mathbb{E}\big[n_t^{(i)}\big] &= \sum_{u=1}^{t} \mathbb{P}(i_u = i) \\
&= \sum_{u=1}^{t} \mathbb{P}(i_u = i, \hat{\mu}_{u-1}^{(i)} + \nu_{u-1}^{(i)} > \mu^{(i_*)}) + \sum_{u=1}^{t} \mathbb{P}(i_u = i, \hat{\mu}_{u-1}^{(i)} + \nu_{u-1}^{(i)} \leqslant \mu^{(i_*)}) \\
&\leqslant \sum_{u=1}^{t} \mathbb{P}\big(i_u = i, \hat{\mu}_{u-1}^{(i)} + \nu_{u-1}^{(i)} > \mu^{(i_*)}\big) + \sum_{u=1}^{t} \mathbb{P}\big(\hat{\mu}_{u-1}^{(i_*)} + \nu_{u-1}^{(i_*)} \leqslant \mu^{(i_*)}\big), \quad (11.15)
\end{aligned}$$

since if we select arm i at time u (i.e., $i_u = i$), then $\hat{\mu}_{u-1}^{(i_*)} + \nu_{u-1}^{(i_*)} \leqslant \hat{\mu}_{u-1}^{(i)} + \nu_{u-1}^{(i)}$.

In order to bound $\mathbb{P}\big(\hat{\mu}_{u-1}^{(i_*)} + \nu_{u-1}^{(i_*)} \leqslant \mu^{(i_*)}\big)$, it is tempting to apply a concentration inequality for the average of $n_{u-1}^{(i_*)}$ independent random variables distributed from the optimal arm distribution. However, these variables are not independent because the obtained reward and the choice of arms are not independent. Instead, we need to use our sequence of i.i.d. samples $x_a^{(i_*)}$, $a \geqslant 1$, with mean $\mu^{(i_*)}$, and bound the probability that *at least* one of these $u-1$ averages of i.i.d. random variables is less than the desired bound. Thus, we have, from sub-Gaussian tail bounds, for $u \in \{1, \dots, t\}$:

$$\mathbb{P}\big(\hat{\mu}_{u-1}^{(i_*)} + \nu_{u-1}^{(i_*)} \leqslant \mu^{(i_*)}\big) \leqslant \sum_{s=1}^{u-1} \exp(-\rho \log(s)) \leqslant \frac{1}{u^{\rho-1}} \leqslant \frac{1}{t^{\rho-1}}.$$

Thus, the right term in Eq. (11.15) is less than $t^{2-\rho}$.

We now bound the left term in Eq. (11.15), as follows, for $t \geqslant k$:

$$\begin{aligned}
&\sum_{u=1}^{t} \mathbb{P}\big(i_u = i, \hat{\mu}_{u-1}^{(i)} + \nu_{u-1}^{(i)} > \mu^{(i_*)}\big) \\
= \;& \sum_{u=1}^{t} \mathbb{P}\Big(i_u = i, \hat{\mu}_{u-1}^{(i)} + \sqrt{2\rho\sigma^2 \log(u)/n_{u-1}^{(i)}} > \mu^{(i_*)}\Big) \\
\leqslant \;& \sum_{u=1}^{t} \mathbb{P}\Big(i_u = i, \hat{\mu}_{u-1}^{(i)} + \sqrt{2\rho\sigma^2 \log(t)/n_{u-1}^{(i)}} > \mu^{(i_*)}\Big) \text{ since } u \leqslant t, \\
= \;& \sum_{u=1}^{t} \sum_{s=1}^{u-1} \mathbb{P}\Big(i_u = i, n_{u-1}^{(i)} = s, \frac{1}{s}\sum_{a=1}^{s} x_a^{(i)} + \sqrt{2\rho\sigma^2 \log(t)/s} > \mu^{(i_*)}\Big).
\end{aligned}$$

We can now swap the two summations to get the bound

$$\begin{aligned}
&\sum_{s=1}^{t-1} \sum_{u=s+1}^{t} \mathbb{P}\Big(i_u = i, n_{u-1}^{(i)} = s, \frac{1}{s}\sum_{a=1}^{s} x_a^{(i)} + \sqrt{2\rho\sigma^2 \log(t)/s} > \mu^{(i_*)}\Big) \\
\leqslant \;& \sum_{s=1}^{t-1} \mathbb{P}\Big(\frac{1}{s}\sum_{a=1}^{s} x_a^{(i)} + \sqrt{2\rho\sigma^2 \log(t)/s} > \mu^{(i_*)}\Big)
\end{aligned}$$

because the events $\{i_u = i, n_{u-1}^{(i)} = s\}$, for $u \in \{s+1, \dots, t\}$ are mutually exclusive. We can then use sub-Gaussian tail bounds, which are non-trivial (that is, less than 1) as soon as $\Delta^{(i)} \geqslant \sqrt{2\rho\sigma^2 \log(t)/s}$, leading to a bound

$$\sum_{s=1}^{+\infty} \exp\big[-\big(\Delta^{(i)} - \sqrt{2\rho\sigma^2 \log(t)/s}\big)_+^2 s/(2\sigma^2)\big].$$

When $s \geqslant \frac{8\rho\sigma^2 \log(t)}{(\Delta^{(i)})^2}$, then the summand is less than $\exp\big[-\frac{s}{4}(\Delta^{(i)})^2/(2\sigma^2)\big]$, and that part of the sum is less than, with $\kappa = \frac{1}{4}(\Delta^{(i)})^2/(2\sigma^2)$:

$$\sum_{s=1}^{\infty} \exp(-s\kappa) = \frac{e^{-\kappa}}{1 - e^{-\kappa}} = \frac{1}{e^{\kappa} - 1} \leqslant \frac{1}{\kappa} = \frac{8\sigma^2}{(\Delta^{(i)})^2}.$$

Otherwise, we bound the probability by one, and get a term equal to $\frac{8\rho\sigma^2 \log(t)}{(\Delta^{(i)})^2}$. Thus, overall we get that

$$\mathbb{E}\big[n_t^{(i)}\big] \leqslant t^{2-\rho} + \frac{8\rho\sigma^2 \log(t)}{(\Delta^{(i)})^2} + \frac{8\sigma^2}{(\Delta^{(i)})^2}.$$

For $\rho = 2$, this leads to a regret bound

$$R_t \leqslant \sum_{i \neq i_*} \Delta^{(i)} \Big(1 + \frac{\sigma^2}{(\Delta^{(i)})^2}(16 \log t + 8)\Big),$$

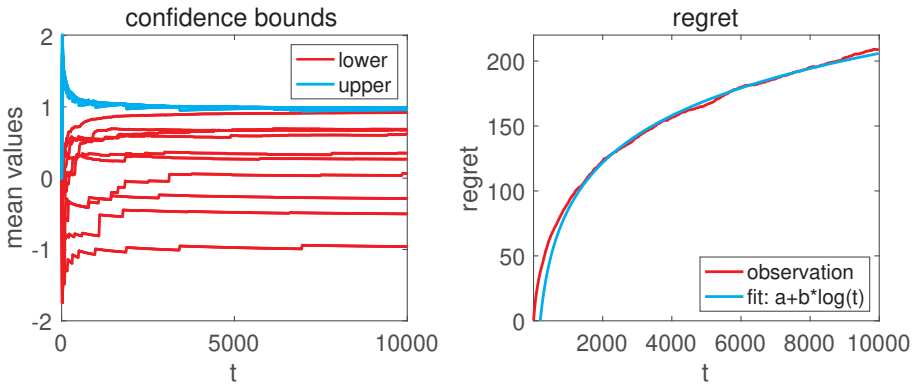


Figure 11.2: Upper-confidence bounds for $k = 10$ Bernoulli arms with random means: plot of upper and lower bounds as a function of time t (left), and regret (right).

which happens to achieve the lower bound (up to constants, see below). We can also obtain a regret that does not blow up when $\Delta^{(i)}$ goes to zero. Indeed, we always have $\sum_{i=1}^{k} n_t^{(i)} \leqslant t$, leading to

$$\begin{aligned} R_t &= \sum_{i,\ \Delta^{(i)}<\Delta} \Delta^{(i)} \mathbb{E}[n_t^{(i)}] + \sum_{i,\ \Delta^{(i)}\geqslant\Delta} \Delta^{(i)} \mathbb{E}[n_t^{(i)}] \text{ for a certain } \Delta, \\ &\leqslant t\Delta + \sum_{i,\ \Delta^{(i)}\geqslant\Delta} \Delta^{(i)} \Big(1 + \frac{\sigma^2}{(\Delta^{(i)})^2}(16 \log t + 8)\Big) \\ &\leqslant t\Delta + \sum_i \Delta^{(i)} + k\frac{\sigma^2}{\Delta}(16 \log t + 8) \leqslant \sum_i \Delta^{(i)} + 8\sigma\sqrt{kt(\log t + 2)}, \end{aligned}$$

by optimizing over Δ, which is also optimal up to logarithmic terms (see below).

If $\rho \geqslant 2$, then we only pay an increase in the bound proportional to ρ, while if $\rho < 2$, the upper-bound on regret can start to be super-linear (and thus vacuous).

Lower bounds. It turns out that with k arms, the best that can be achieved is a regret of order $\sigma\sqrt{kt}$, and for the instance-dependent problem, or order $\log(t) \sum_{i \neq i_*} \frac{\sigma^2}{\Delta^{(i)}}$ (see, e.g., Bubeck and Cesa-Bianchi, 2012).

Illustration. In Figure 11.2, we plot the performance of the UCB algorithm with $k = 10$ arms. In particular, the right plot highlights that the upper confidence bounds for all arms tend to be equal.

11.3.4 Adversarial bandits (♦)

We finish this section on multi-armed bandits by studying a non-stochastic set-up referred to as the adversarial set-up. We now have arbitrary reward vectors $\mu_t \in [0, 1]^k$, $t \geqslant 1$, that

may vary with time and are assumed deterministic, and at each time step, we choose an arm i_t, and we receive reward $\mu_t^{(i_t)}$. This context where the reward vectors are selected in advance (but arbitrary and unknown) is referred to as an "oblivious" adversary, as opposed to an "adaptive" adversary where the functions can depend on past information.

The regret is then

$$\max_{i\in\{1,\dots,k\}} \sum_{s=1}^{t} \mu_s^{(i)} - \sum_{s=1}^{t} \mu_s^{(i_s)}.$$

Note that in this set-up, there is no randomness in the environment and that we receive rewards that are elements of $[0,1]$ and not sampled in $\{0,1\}$ from that quantity. The stochastic setting can be seen as a particular subcase (but for which other algorithms, such as UCB, can be applied; see a comparison at the end of this section).

Impossibility of deterministic policies. If the choice of $i_t \in \{1,\dots,k\}$ is deterministic (and function of the past information), then there exists a reward sequence (μ_t) so that $\mu_t^{(i_t)} = 0$ and $\mu_t^{(i)} = 1$ for $i \neq i_t$. After t steps, at least one arm has been chosen less than t/k times. For that arm $\sum_{s=1}^{t} \mu_s^{(i)} \geqslant t - t/k$, and thus the regret is greater than $t(1-1/k)$ which is linear in t.

We, therefore, consider expectations from a randomized algorithm.

Hedge algorithm (♦). We start with the situation where a full reward vector $\mu_t \in [0,1]^k$ is observed at every iteration (after the choice of action, which is sampled from the probability vector π_{t-1} in the simplex in k dimensions). The expectation of the reward is then $\pi_{t-1}^\top \mu_t$. The Hedge algorithm (Freund and Schapire, 1997) consists in starting with π_0 uniform and updating π_t as follows:

$$\forall i \in \{1,\dots,n\},\ \pi_t^{(i)} = \frac{\pi_{t-1}^{(i)} \exp(\gamma \mu_t^{(i)})}{\sum_{j=1}^{k} \pi_{t-1}^{(j)} \exp(\gamma \mu_t^{(j)})},$$

where $\gamma > 0$ is a free parameter. This happens to be exactly the online mirror descent algorithm from Section 11.1.3, with no randomness, applied to $F_t(\pi) = \mu_t^\top \pi$, with the entropy mirror map. We thus get immediately an expected normalized regret (with respect to the randomization of the algorithm) less than $\sqrt{2\log(k)}/\sqrt{t}$ for the choice $\gamma = \sqrt{2\log(k)}/\sqrt{t}$. We therefore get a (unnormalized) regret proportional to $\sqrt{t\log(k)}$.

Exp3 algorithm (♦♦). To tackle the bandit case with limited feedback, we follow the same strategy as the Hedge algorithm but with an unbiased estimator of the vector $\mu_t \in [0,1]^k$, from which we only observe the component $\mu_t^{(i_t)}$, where i_t is sampled from π_{t-1}. The estimator suggested by Auer et al. (2002b) is an importance sampling estimator and leads to the "Exp3" algorithm. It is defined as $\hat{\mu}_t^{(i)} = \mu_t^{(i_t)} 1_{i=i_t} / \pi_{t-1}^{(i)}$; it thus has expectation μ_t, and variance $\mathbb{E}\big[\|\hat{\mu}_t\|_\infty^2\big] \leqslant \mathbb{E}\big[\|\hat{\mu}_t\|_2^2\big] \leqslant \sum_{i=1}^{k} 1/\pi_{t-1}^{(i)}$, which is not enough to get a non-explosive bound. However, an improvement on Prop. 11.1 may be obtained for the simplex.

Proposition 11.3 *The mirror descent recursion in Eq. (11.3), for $\mathcal{C}$ the simplex and Φ the entropy mirror map, is equal to $\theta_t^{(i)} = \theta_{t-1}^{(i)} \exp(-\gamma g_t^{(i)}) / \sum_{j=1}^d \theta_{t-1}^{(j)} \exp(-\gamma g_t^{(j)})$ for all $i \in \{1, \dots, d\}$. Then, assuming that g_t has almost surely non-negative components and $\mathbb{E}[\sum_{i=1}^d \theta_{s-1}^{(i)} (g_s^{(i)})^2 | \mathcal{F}_{s-1}] \leqslant B^2$ almost surely for all $s \geqslant 1$, for every $\theta \in \mathcal{C}$, we have:*

$$\frac{1}{t} \sum_{s=1}^t \mathbb{E}\big[F_s(\theta_{s-1}) - F_s(\theta)\big] \leqslant \frac{1}{\gamma t} D_\Phi(\theta, \theta_0) + \frac{\gamma B^2}{2}.$$

Proof Following the proof of Prop. 11.1, we have $D_\Phi(\theta_t, \theta_{t-1}) + \gamma g_t^\top (\theta_t - \theta_{t-1}) = -\gamma \sum_{i=1}^d \theta_{t-1}^{(i)} g_t^{(i)} - \log \sum_{i=1}^d \exp(-\gamma g_t^{(i)})$, which, using Exercise 1.14, is greater than $-\frac{\gamma^2}{2} \sum_{i=1}^d \theta_{t-1}^{(i)} (g_t^{(i)})^2$. This leads to the desired result. ∎

We can now provide a regret bound for the Exp3 algorithm by using Prop. 11.3 and noticing that we need to bound $\mathbb{E}[\sum_{i=1}^d \pi_{s-1}^{(i)} (\hat{\mu}_s^{(i)})^2 | \mathcal{F}_{s-1}] = \sum_{i=1}^d (\mu_t^{(i)})^2 \leqslant k$. This leads to, after optimizing with respect to the step-size, a non-normalized regret bound proportional to $\sqrt{kt \log k}$.

From adversarial to stochastic. In the stochastic set-up, the UCB algorithms provided an un-normalized regret of order $\sqrt{kt \log k}$, which is the same as the regret of the Exp3 algorithm, which is aimed at the adversarial setting. For an analysis of Exp3 in the stochastic case, see Seldin et al. (2013) for a similar regret bound.

11.4 Conclusion

In this chapter, we have provided extensions to the classical independent and identically distributed setting that is the book's main focus. In the convex case, algorithms and analysis were similar to the classical case, and seamlessly allowed arbitrary sequences of functions to be optimized. In the bandit setting, where only partial information was provided, a dedicated algorithmic framework was presented (optimism in front of uncertainty). There are multiple extensions, as described byShalev-Shwartz (2011); Bubeck and Cesa-Bianchi (2012); Hazan (2022); Slivkins (2019); Lattimore and Szepesvári (2020).