Algorithmic Foundations of Learning

Problem Sheet 4

Lecturer: Patrick Rebeschini *Due date: 48 hours before class*

4.1 High-Probability Condition for Restricted Strong Convexity (Question type: B)

Prove Proposition 13.6 in the Lecture Notes.

Hint: Look at $\mathbf{P}(\|\frac{\mathbf{X}^\top \mathbf{X}}{n} - I\| \geq \varepsilon)$ and use the union bound.

4.2 Full Information Setting and Bounded Pseudo-Regret Policy (Question type: B)

Consider the following full information online statistical learning problem with $k = 2$ actions and time horizon $n > 0$. At every time step $t = 1, 2, \ldots, n$:

1. Choose an action $A_t \in \{1, 2\}$;
2. Observe a reward vector $Z_t = (Z_{t,1}, Z_{t,2}) \in \mathbb{R}^2$;
3. Suffer a loss $\ell(A_t, Z_t) = -Z_{t,A_t}$.

Consider the policy given by $A_1 = 1$, $A_2 = 2$, and, for any $t \geq 3$,

$$A_t \in \operatorname*{argmax}_{a \in \{1,2\}} M_{t-1,a},$$

where $M_{t,a} := \frac{1}{t}\sum_{s=1}^{t} Z_{s,a}$ for any $a \in \{1, 2\}$. Assume that the two components of the reward vector are sampled independently from two sub-Gaussian distributions with the same variance proxy σ^2. Prove that the policy incurs a pseudo-regret R_n that can be bounded as follows:

$$\mathbf{E}R_n \leq \Delta + \frac{4\sigma^2}{\Delta},$$

where Δ is the sub-optimality gap.

Hint: Write the total number of times the sub-optimal arm is played in terms of the sample means $M_{t,a}$, for $a \in \{1, 2\}$ and $t \in [n]$.

Remark. In the full information setting, a simple strategy based on playing the arm that has achieved the highest sample mean leads to a pseudo-regret bounded by a quantity that does not depend on the time horizon n.

4.3 Properties of the KL Divergence and Pinsker's inequality (Question type: A)

Prove Proposition 16.4 in the Lecture Notes.

Hint: To prove the Gibbs' inequality, use that the function $x \in \mathbb{R}_+ \to f(x) := x \log x$ is convex. To prove Pinsker's inequality, first prove that if X is a random variable distributed according to p and E is a measurable event, then $\mathbf{E}[\mathbf{1}_E(X) \log \frac{p(X)}{q(X)}] \geq \mathbf{P}(E) \log \frac{\mathbf{P}(E)}{\mathbf{Q}(E)}$. Use this to prove that $\mathtt{KL}(\mathbf{P}, \mathbf{Q}) \geq \mathtt{KL}(\mathrm{Bern}(\mathbf{P}(E)), \mathrm{Bern}(\mathbf{Q}(E)))$, where $\mathrm{Bern}(a)$ is a Bernoulli distribution with parameter a.

4.4 On the Minimal Amount of Information with Coin Flips (Question type: B)

You observe a sequence of n independent coin flips $X_1, \ldots, X_n$ that are generated either by a fair coin (i.e., the coin flips follow a Bernoulli distribution with mean $\mu_0 = 1/2$) or by a biased coin with bias $\varepsilon \in (0, 1/4)$ (i.e., the coin flips follow a Bernoulli distribution with mean $\mu_1 = 1/2 + \varepsilon$). You want to design a test $f : \{0,1\}^n \to \{0,1\}$ that performs better than random guessing in each scenario, namely, that satisfies

$$\mathbf{P}_{\mu_0}(f(X_1, \ldots, X_n) = 0) \geq c,$$
$$\mathbf{P}_{\mu_1}(f(X_1, \ldots, X_n) = 1) \geq c,$$

for a given $c > 1/2$. Show that you need at least $n \geq \frac{(2c-1)^2}{2\varepsilon^2}$ coin flips.

Remark. It is expected that the smaller the bias ε is the more difficult the testing problem becomes. However, note that the difficulty scales *quadratically* with ε, not linearly.

4.5 Variance Reduction for Stochastic Gradient Descent (Question type: B)

Let $f_1, \ldots, f_n$ be β-smooth convex functions from $\mathbb{R}^d$ to $\mathbb{R}$, and let $f := \frac{1}{n}\sum_{i=1}^n f_i$ be α-strongly convex. Let $x^\star$ be the minimizer of f. It can be shown that if I is uniformly distributed in $\{1, \ldots, n\}$, then for any $x \in \mathbb{R}^d$ we have

$$\mathbf{E}\|\nabla f_I(x) - \nabla f_I(x^\star)\|_2^2 \leq 2\beta(f(x) - f(x^\star)).$$

Answer the following questions.

1. For a given $y \in \mathbb{R}^d$ and $\eta \in (0, \frac{1}{2\beta})$, consider the following algorithm:

$$X_1 = y,$$
$$X_{s+1} = X_s - \eta(\nabla f_{I_{s+1}}(X_s) - \nabla f_{I_{s+1}}(y) + \nabla f(y)) \qquad \text{for } s = 1, \ldots, k,$$

where $I_2, \ldots, I_{k+1}$ is a collection of i.i.d. random variables uniformly distributed in $\{1, \ldots, n\}$. Prove that

$$\mathbf{E} f\left(\frac{1}{k}\sum_{s=1}^k X_s\right) - f(x^\star) \leq \left(\frac{1}{\alpha\eta(1-2\beta\eta)k} + \frac{2\beta\eta}{1-2\beta\eta}\right)(f(y) - f(x^\star)).$$

Hint: Compute an upper bound for $\mathbf{E}[\|X_{k+1}-x^\star\|_2^2|X_k]$ by writing $\|X_{k+1}-x^\star\|_2^2$ in terms of $\|X_k-x^\star\|_2^2$ and by using the convexity of f. You might find useful the inequalities: $(a+b)^2 \leq 2a^2+2b^2$ for any $a,b\in\mathbb{R}$; $\mathbf{E}[\|X-\mathbf{E}X\|_2^2] \leq \mathbf{E}[\|X\|_2^2]$ for any random variable X; $f(x)-f(x^\star) \geq \frac{\alpha}{2}\|x-x^\star\|_2^2$ for any $x\in\mathbb{R}^d$, by strong convexity of f as $\nabla f(x^\star)=0$.

2. Describe an algorithm that computes a ε-approximate solution $\widetilde{X}$ to $x^\star$ (i.e., that satisfies $\mathbf{E}f(\widetilde{X}) - f(x^\star) \leq \varepsilon$) with a computational complexity that scales like $O((n+\beta/\alpha)\log(1/\varepsilon))$. In the light of the results discussed in the course, do you find this fact surprising? Why?

4.6 Collaborative Filtering (Question type: C)

You run a business such as Yelp or Netflix, where you want to predict the ratings that d_1 users give to d_2 products. The unknown parameter that you want to learn is the preference matrix $\mathbf{w}^\star \in \mathbb{R}^{d_1\times d_2}$, where $\mathbf{w}^\star_{ij}$ is the rating that user i gives to product j. You have access to n noisy measurements of the entries of $\mathbf{w}^\star$, corresponding to the ratings that users gave to products they have used (e.g., restaurants they have been to in the Yelp example, or movies they have seen in the Netflix example), corrupted by noise. The noise assumption is meant to take into account inaccurate ratings due, say, to users having a good or a bad day when they rate a product. You are dealing with a high-dimensional set up, where the number of observations at your disposal is much less than the number of parameters you want to infer: $n \ll d_1 \times d_2$. Observations are represented as pairs $\{\mathbf{x}_\ell, Y_\ell\}$, $\ell\in[n]$. Here, $\mathbf{x}_\ell = \mathbf{1}_{i(\ell)}\mathbf{1}_{j(\ell)}^\top \in \mathbb{R}^{d_1\times d_2}$, where $i(\ell)$ and $j(\ell)$ are, respectively, the user and product associated to the ℓ-th observation. You consider the following model:

$$Y_\ell = \langle \mathbf{x}_\ell, \mathbf{w}^\star\rangle + \xi_\ell \in \mathbb{R},$$

where the noise parameter ξ_ℓ is an independent standard Gaussian random variable, and where $\langle\,\cdot\,,\,\cdot\,\rangle$ denotes the matrix inner product defined as $\langle \mathbf{a},\mathbf{b}\rangle := \mathtt{Trace}(\mathbf{a}^\top\mathbf{b})$. This is a component-wise inner product that corresponds to the standard inner product of vectors when the entries of a matrix are considered as the components of a corresponding vector.

1. Prove that $Y_\ell = \mathbf{w}^\star_{i(\ell)j(\ell)} + \xi_\ell$.
2. Given the high-dimensional nature of the problem, you need to assume that the unknown parameter lies in a low dimensional space (see Section 12.2 in the Lecture Notes). If you run a business like Yelp, explain why the structure $\mathbf{w}^\star = \mathbf{1}(v^\star)^\top$, for a given vector $v^\star\in\mathbb{R}^{d_2}$, is a reasonable assumption. On the other hand, if you run a business like Netflix, explain why the more general structure $\mathbf{w}^\star = \mathbf{u}^\star(\mathbf{v}^\star)^\top$, where $\mathbf{u}^\star\in\mathbb{R}^{d_1\times k}$ and $\mathbf{v}^\star\in\mathbb{R}^{d_2\times k}$ for a given $k>0$, is needed. In particular, what is k?
3. Consider the estimator
$$\mathbf{W} := \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^{d_1\times d_2}} R(\mathbf{w}) + \lambda\|\mathbf{w}\|,$$
where $R(\mathbf{w}) := \frac{1}{2n}\sum_{\ell=1}^n(\langle\mathbf{x}_\ell,\mathbf{w}\rangle - Y_\ell)^2$ and where $\|\cdot\|$ represents the nuclear norm of a matrix defined as $\|\mathbf{a}\| := \sum_{i=1}^{\min\{d_1,d_2\}}\sigma_i(\mathbf{a})$, where $\sigma_i(\mathbf{a})$ are the singular values of the matrix $\mathbf{a}$. Show that the dual of the nuclear norm is given by $\|\mathbf{a}\|_* = \max_{i\in\min\{d_1,d_2\}}\sigma_i(\mathbf{a})$.
4. For a given $\mathbf{a}\in\mathbb{R}^{d_1\times d_2}$ and a subspace $\mathcal{S}\subseteq\mathbb{R}^{d_1\times d_2}$, denote by $\mathbf{a}_\mathcal{S}$ the projection of $\mathbf{a}$ into $\mathcal{S}$ defined as
$$\mathbf{a}_\mathcal{S} := \operatorname*{argmin}_{\mathbf{b}\in\mathcal{S}} \|\mathbf{a}-\mathbf{b}\|_F,$$
where $\|\cdot\|_F$ is the Frobenius norm for matrices defined as $\|\mathbf{a}\|_F := \sqrt{\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\mathbf{a}_{ij}^2} = \sqrt{\mathtt{Trace}(\mathbf{a}^\top\mathbf{a})}$. Denote by $\mathcal{S}^\perp$ the orthogonal complement of $\mathcal{S}$, defined as
$$\mathcal{S}^\perp := \{\mathbf{a}\in\mathbb{R}^{d_1\times d_2} : \langle\mathbf{a},\mathbf{b}\rangle = 0 \text{ for all } \mathbf{b}\in\mathcal{S}\}.$$

Assume that $\mathbf{w}^\star = \mathbf{u}^\star(\mathbf{v}^\star)^\top$, where $\mathbf{u}^\star \in \mathbb{R}^{d_1 \times k}$ and $\mathbf{v}^\star \in \mathbb{R}^{d_2 \times k}$ for a given $k > 0$, and define:

$$\mathcal{M} := \{\mathbf{w} \in \mathbb{R}^{d_1 \times d_2} : \texttt{Span}(\mathbf{w}^\top) \subseteq \texttt{Span}(\mathbf{v}^\star), \texttt{Span}(\mathbf{w}) \subseteq \texttt{Span}(\mathbf{u}^\star)\},$$
$$\overline{\mathcal{M}}^\perp := \{\mathbf{w} \in \mathbb{R}^{d_1 \times d_2} : \texttt{Span}(\mathbf{w}^\top) \subseteq \texttt{Span}(\mathbf{v}^\star)^\perp, \texttt{Span}(\mathbf{w}) \subseteq \texttt{Span}(\mathbf{u}^\star)^\perp\},$$

where $\texttt{Span}(\mathbf{a})$ is the subspace obtained by the linear combinations of the column vectors of the matrix $\mathbf{a}$. Note that $\mathbf{w}^\star \in \mathcal{M}$. It can be shown that if $\mathbf{a} \in \mathcal{M}$ and $\mathbf{b} \in \overline{\mathcal{M}}^\perp$ then the nuclear norm decomposes, namely, $\|a + b\| = \|a\| + \|b\|$.

Consider the following assumption.

Assumption 4.1 *Define the cone set*

$$\boxed{\mathcal{C} := \{\mathbf{w} \in \mathbb{R}^{d_1 \times d_2} : \|\mathbf{w}_{\overline{\mathcal{M}}^\perp}\| \leq 3\|\mathbf{w}_{\overline{\mathcal{M}}}\|\}}$$

There exists $\alpha > 0$ such that

$$\boxed{R(\mathbf{w}^\star + \mathbf{w}) \geq R(\mathbf{w}^\star) + \langle \nabla R(\mathbf{w}^\star), \mathbf{w}\rangle + \alpha\|\mathbf{w}\|_F^2 \qquad \textit{for any } w \in \mathcal{C}} \tag{4.1}$$

Prove the following theorem.

Theorem 4.2 *If Assumption 4.1 holds and $\lambda \geq 2\|\nabla R(\mathbf{w}^\star)\|_*$, then*

$$\boxed{\|\mathbf{W} - \mathbf{w}^\star\|_F \leq \frac{3\lambda}{2\alpha}\Psi(\overline{\mathcal{M}})}$$

where $\Psi(\overline{\mathcal{M}}) := \sup_{\mathbf{a} \in \overline{\mathcal{M}} \setminus \{\mathbf{0}\}} \frac{\|a\|}{\|a\|_F}$.

Hint: Follow the proof of Theorem 13.4 in the Lecture Notes, use the decompositions $\mathbf{w}^\star = \mathbf{w}^\star_{\mathcal{M}} + \mathbf{w}^\star_{\mathcal{M}^\perp}$ and $\Delta = \Delta_{\overline{\mathcal{M}}} + \Delta_{\overline{\mathcal{M}}^\perp}$ and that, by the reverse triangle inequality, $\|\mathbf{w}^\star + \Delta\| \geq \|\mathbf{w}^\star_{\mathcal{M}} + \Delta_{\overline{\mathcal{M}}^\perp}\| - \|\mathbf{w}^\star_{\mathcal{M}^\perp}\| - \|\Delta_{\overline{\mathcal{M}}}\|$.