Stochastic Foundations of Decision-Aware Learning **Mini-Course 2022**

Lecture 1: Proxy-Objectives and Decision-Aware Learning

Instructor: Vishal Gupta guptavis@usc.edu

1. Course Overview

"Decision-aware" learning (also known as optimization-aware learning, end-to-end learning, task-based learning, decision-focused learning, and operational statistics) is an increasingly popular research area within data-driven optimization. Indeed, empirical studies of "decision-aware" algorithms often show they can substantially outperform traditional "predict-then-optimize" approaches to decision-making under uncertainty, but a complete theoretical understanding of their strengths and weaknesses is still developing.

This mini-course is meant as a quick & dirty introduction to the use of proxy-objectives for decision-aware learning, and analyzing those proxy-objectives via empirical process theory. Our focus is on how to use the theory to prove performance guarantees about various methods. The aim is to empower students to use these methods in their own research.

2. Warm-Up: Sample Average Approximation (SAA)

Traditional Stochastic Optimization studies:

$$\boldsymbol{x}^* \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}\left[c(\boldsymbol{x},\boldsymbol{\xi})\right], \tag{1-1}$$

where

- $\boldsymbol{x}$ is our decision variable,
- $\mathcal{X}$ is a *known* feasible region,
- $c(\cdot,\cdot)$ is a *known* cost function,
- and $\boldsymbol{\xi} \sim \mathbb{P}$ is an *exogenous* random variable.

Exercise: Describe some applications of stochastic optimization as described below. Be specific. What do each of $\boldsymbol{x}$, $\boldsymbol{\xi}$, $c(\cdot,\cdot)$ and $\mathcal{X}$ model?

Importantly, in the data-driven setting, $\mathbb{P}$ is *unknown*, so we cannot compute $\boldsymbol{x}^*$ directly. Rather, we have access to data $\{\hat{\boldsymbol{\xi}}_1,\ldots,\hat{\boldsymbol{\xi}}_n\}$ drawn i.i.d. from $\mathbb{P}$.

With these data, we might use the Sample Average Approximation (SAA) (also called Empirical Risk Minimization or ERM):

$$\boldsymbol{x}_n \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\sum_{j=1}^{n} c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j). \tag{1-2}$$

How should we assess the quality of $\boldsymbol{x}_n$? Since $\boldsymbol{x}_n$ is feasible in Problem (1-1) by construction, we often just upper bound its out-of-sample performance $\mathbb{E}\left[c(\boldsymbol{x}_n, \boldsymbol{\xi}) \mid \hat{\boldsymbol{\xi}}_{1:n}\right]$.

IMPORTANT: *Out-of-sample performance is a random variable. (Why?) Hence, an upper bound is usually bounds its expectation or (better) its tail behavior.*

This upper bound is usually computed relative to a suitable oracle benchmark. For example, in a few sessions, we will prove the following theorem:

Theorem 1.1 (SAA for Discrete Feasible Regions)**.** *Suppose* $|\mathcal{X}| < \infty$*, and* $\max_{\boldsymbol{x}\in\mathcal{X}} |c(\boldsymbol{x}, \boldsymbol{\xi})| \leq c_{\max}$ *almost surely. Then there exists a universal constant* $C > 0$ *such that for any* $0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$0 \leq \mathbb{E}\left[c(\boldsymbol{x}_n, \boldsymbol{\xi}) \mid \hat{\boldsymbol{\xi}}_{1:n}\right] - \mathbb{E}\left[c(\boldsymbol{x}^*, \boldsymbol{\xi})\right] \leq C \cdot c_{\max} \sqrt{\frac{\log(2/\delta)\log(|\mathcal{X}|)}{n}}.$$

Here the "oracle benchmark" is the performance of $\boldsymbol{x}^*$, i.e., the performance we would have gotten had we known $\mathbb{P}$ upfront. Thus, in words, the theorem says that with high probability, the out-of-sample performance is not much worse than the best possible performance, provided n is much larger that $\log(|\mathcal{X}|)$. Note that we can provide this bound, even though we *can't* evaluate $\boldsymbol{x}^*$!

3. The Proxy-Objective Perspective

SAA is our first example of the use of a proxy-objective in decision-aware learning. Specifically,

- We would like to solve Problem (1-1) but we cannot because we don't know $\mathbb{P}$, and thus don't know the objective function.
- Instead, we solve a different optimization problem (i.e., Problem (1-2)) with an objective function built from the data and meant to approximate the unknown objective function.
- We provide a performance guarantee on the suboptimality of our proxy problem's solution in the original problem.

Many methods in data-driven optimization follow this loose structure and, consequently, can be analyzed using a small set of similar tools. We next give some other examples.

3.1. Parametric Probability Models.

Suppose we believe the unknown $\mathbb{P}$ follows some parametric model, e.g., that it is a multivariate normal distribution with known covariance equal to the identity.

Then, leting $\mathbb{P} = \mathbb{P}_{\boldsymbol{\theta}^*}$ for some parameters $\boldsymbol{\theta}^* \in \Theta$, we rewrite Problem (1-1) as

$$\boldsymbol{x}(\boldsymbol{\theta}^*) \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}^{\mathbb{P}_{\boldsymbol{\theta}^*}}\left[c(\boldsymbol{x}, \boldsymbol{\xi})\right]. \tag{1-3}$$

We might then use our favorite statistical learning procedure (e.g. Maximum Likelihood) to form an estimate $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\hat{\boldsymbol{\xi}}_{1:n})$ from the data and compute a predict-then-optimize decision:

$$\boldsymbol{x}(\hat{\boldsymbol{\theta}}_n) \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}^{\mathbb{P}_{\hat{\boldsymbol{\theta}}_n}}\left[c(\boldsymbol{x}, \boldsymbol{\xi})\right] \tag{1-4}$$

In other words, we use Problem (1-4) as a proxy for Problem (1-3). We might then seek high probability bounds on

$$\mathbb{E}\left[c(\boldsymbol{x}(\hat{\boldsymbol{\theta}}_n), \boldsymbol{\xi}) \mid \hat{\boldsymbol{\xi}}_{1:n}\right] - \mathbb{E}\left[c(\boldsymbol{x}(\boldsymbol{\theta}^*), \boldsymbol{\xi})\right]$$

to establish a performance guarantee.

3.2. **Policy Classes and Best-in-Class Policies.** What happens if n if the same order as $\log(|\mathcal{X}|)$in Theorem 1.1? Our performance guarantee is not informative since it's proportional $c_{\max}$ and $2c_{\max}$ is a trivial performance guarantee.

This observation isn't a weakness in our analysis.

Exercise: Construct an example of binary optimization of the form Problem (1-1) where the performance of SAA is (with high probability) far from the full-information optimum when the number of data points is equal to the number of binary variables.

In fact, in certain data settings, e.g., when data are scarce, *every* method has a large sub-optimality gap relative to the full-information optimal solution.

So what do we do in these cases? Since, full-information isn't achievable, it makes sense to consider a weaker oracle benchmark. Inspired by Theorem 1.1, one approach might be to identify a subset $\mathcal{X}_0 \subseteq \mathcal{X}$ of the feasible region, and limit attention to solutions in $\mathcal{X}_0$.

Following this line of reasoning, we define

$$\boldsymbol{x}_0^* = \boldsymbol{x}_0^*(\boldsymbol{\xi}_{1:n}) \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{X}_0} \mathbb{E}\left[c(\boldsymbol{x}, \boldsymbol{\xi})\right] \tag{1-5}$$

to be the best solution in $\mathcal{X}_0$. The "best-in-class" solution cannot be computed because $\mathbb{P}$ is unknown, but performs (path by path) every other policy with values in $\mathcal{X}_0$.

Since we can't solve Problem (1-5), we proxy it by

$$\hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{x}}_0(\hat{\boldsymbol{\xi}}_{1:n}) \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{X}_0} \frac{1}{n} \sum_{j=1}^{n} c(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_j), \tag{1-6}$$

i.e., the "corresponding" SAA solution. Theorem 1.1 then gives a performance guarantee: with probability at least $1 - \delta$,

$$0 \leq \mathbb{E}\left[c(\hat{\boldsymbol{x}}_0, \boldsymbol{\xi}) \mid \hat{\boldsymbol{\xi}}_{1:n}\right] - \mathbb{E}\left[c(\boldsymbol{x}_0^*, \boldsymbol{\xi})\right] \leq C \cdot c_{\max} \sqrt{\frac{\log(2/\delta)\log(|\mathcal{X}_0|}{n}}. \tag{1-7}$$

Thus, if $\log(|\mathcal{X}_0|) \ll n$, we can learn a solution $\hat{\boldsymbol{x}}_0$ whose performance is comparable to the best-in-class solution $\boldsymbol{x}_0$.

Of course, there are no free lunches. The above result tells us nothing about the gap $\mathbb{E}\left[c(\boldsymbol{x}_0, \boldsymbol{\xi})\right] - \mathbb{E}\left[c(\boldsymbol{x}^*, \boldsymbol{\xi})\right]$. If $|\mathcal{X}_0|$ is small, we might expect this gap to be big. But, if $|\mathcal{X}_0|$ is big, then the performance guarantee Eq. (1-7) will be poor and our previous exercise suggests it may be hard to learn a best-in-class policy. This tension is somehow unavoidable.

In summary, in settings where the data are rich/plentiful enough, we should try to learn a policy with performance similar to the full-information solution $\boldsymbol{x}^*$. In setting where this is not possible, we should instead use our domain knowledge to restrict attention to a smaller set of policies, and

try to learn a policy with performance comparable to the best policy in that set. There are of course *many* ways to choose the "smaller set of policies," and a good choice is typically application dependent.

For now, what I want to stress it that this example again follows our proxy-objective framework. We proxy Problem (1-5) by Problem (1-6) and prove a performance guarantee on the gap to best-in-class.

Remark 1.2 (Error from Policy Restriction)**.** *After class, a smart student asked about bounding the gap* $\mathbb{E}\left[c(\boldsymbol{x}_0^*, \boldsymbol{\xi})\right] - \mathbb{E}\left[c(\boldsymbol{x}^*, \boldsymbol{\xi})\right]$*. In some very special situations, it is possible to bound this error, however, this is usually the exception, not the rule. In this course, we will focus on the cases where* $\mathcal{X}_0$ *is determined by domain knowledge and practical considerations, and focus simply on learning a best-in-class policy.*

3.3. **Plug-In Policy Classes.** Above, the set $\mathcal{X}_0$ does not depend on the data. (But, both the policies $\boldsymbol{x}_0^*$ and $\hat{\boldsymbol{x}}_0$ *do* depend on the data. Why?) Identifying a good $\mathcal{X}_0$ so that $|\mathcal{X}_0|$ is small but still sufficiently rich so that $\mathbb{E}\left[c(\boldsymbol{x}_0, \boldsymbol{\xi})\right] - \mathbb{E}\left[c(\boldsymbol{x}^*, \boldsymbol{\xi})\right]$ is small from just prior domain knowledge might be tricky.

We next introduce a particular class of policies called "plug-in policies" that take values in a set $\mathcal{X}_0(\hat{\boldsymbol{\xi}}_{1:n})$ which does depend on the data. These policies have some nice practical features. For example, plug-in policies generalize the popular "estimate-then-optimize" approach to a decision-aware setting. Some authors prefer the term "smart predict-then-optimize" to describe best-in-class plug-in policies [EG22].

To that end, rewrite the full-information optimal solution to Problem (1-1) as

$$\boldsymbol{x}(\mathbb{P}) \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}_{\boldsymbol{\xi} \sim \mathbb{P}}\left[c(\boldsymbol{x}, \boldsymbol{\xi})\right],$$

to stress the dependence on $\mathbb{P}$. That is, $\boldsymbol{x}(\mathbb{P})$ is the optimal solution when the data follows distribution $\mathbb{P}$.

More generally, when $\boldsymbol{\xi} \sim \mathbb{Q}$, define the plug-in policy for $\mathbb{Q}$,

$$\boldsymbol{x}(\mathbb{Q}) \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}_{\boldsymbol{\xi} \sim \mathbb{Q}}\left[c(\boldsymbol{x}, \boldsymbol{\xi})\right]. \tag{1-8}$$

These policies essentially "plug-in" $\mathbb{Q}$ for the unknown $\mathbb{P}$ and solve the optimization problem.

We can now define a class of plug-in policies. Consider a set of probability distributions (dependent on data) $\{\hat{\mathbb{P}}_\tau(\hat{\boldsymbol{\xi}}_{1:n}) : \tau \in \mathcal{T}\}$, where for each τ, $\hat{P}_\tau(\hat{\boldsymbol{\xi}}_{1:n})$ refers to a different way to fit a probability distribution to the observed data For example, one τ might correspond to fitting a mixture of normal distributions to the data, while another might correspond to simply using the empirical distribution.

This set induces a set of (data-dependent) plug-in policies

$$\mathcal{X}_0(\hat{\boldsymbol{\xi}}_{1:n}) \;=\; \left\{\boldsymbol{x}(\mathbb{Q}) : \mathbb{Q} \in \{\hat{\mathbb{P}}_\tau(\hat{\boldsymbol{\xi}}_{1:n}) : \tau \in \mathcal{T}\}\right\} \;=\; \left\{\boldsymbol{x}(\mathbb{P}_\tau(\hat{\boldsymbol{\xi}}_{1:n})) : \tau \in \mathcal{T}\right\}.$$

Both the policy class $\mathcal{X}_0(\hat{\boldsymbol{\xi}}_{1:n})$ *and* each plug-in policy $\boldsymbol{x}(\mathbb{P}_\tau(\hat{\boldsymbol{\xi}}_{1:n})) \in \mathcal{X}_0(\hat{\boldsymbol{\xi}}_{1:n})$ depend on the data.

We define the (oracle) best-in-class decision $\boldsymbol{x}(\mathbb{P}_{\tau^{\mathrm{OR}}}(\hat{\boldsymbol{\xi}}_{1:n}))$, where

$$\tau^{\mathrm{OR}} \in \operatorname*{argmin}_{\tau\in\mathcal{T}} \mathbb{E}\left[c(\boldsymbol{x}(\mathbb{P}_\tau(\hat{\boldsymbol{\xi}}_{1:n})),\boldsymbol{\xi}) \mid \hat{\boldsymbol{\xi}}_{1:n}\right]. \tag{1-9}$$

By construction, this oracle benchmark outperforms any other plug-in policy in $\mathcal{X}_0(\hat{\boldsymbol{\xi}}_{1:n})$. However, it is an "oracle" benchmark because one must know the distribution $\mathbb{P}$ to compute τ^{OR}.

Since we can't solve Problem (1-9) directly, we create a proxy. One natural proxy might be the leave-one-out objective:

$$\tau^{\mathrm{LOO}} \in \operatorname*{argmin}_{\tau\in\mathcal{T}} \sum_{j=1}^{n} c(\boldsymbol{x}(\hat{\mathbb{P}}_\tau(\hat{\boldsymbol{\xi}}_{-j}),\hat{\boldsymbol{\xi}}_j), \tag{1-10}$$

where $\hat{\boldsymbol{\xi}}_{-j}$ is the data $\hat{\boldsymbol{\xi}}_{1:n}$ *excluding* data point j.

Thus, we have proxied Problem (1-9) by Problem (1-10). We might then seek to bound

$$\mathbb{E}\left[c(\boldsymbol{x}(\hat{\boldsymbol{\xi}}_{1:n},\tau^{\mathrm{LOO}}),\boldsymbol{\xi})\right] - \mathbb{E}\left[c(\boldsymbol{x}(\hat{\boldsymbol{\xi}}_{1:n},\tau^{\mathrm{OR}}),\boldsymbol{\xi})\right]$$

to provide a performance guarantee of our method relative to the best-in-class plug-in policy.

The notation in this example is a bit heavy. For now, things I want to stress:

- When considering "best-in-class" learning, we need to create a proxy for the best-in-class optimization problem. By contrast, in our example in Section 2 needed to create a proxy for the full-information problem. You should look to see how this difference also changes the kind of performance guarantee we seek.
- The optimization Problem (1-10) can be quite difficult unless $\mathcal{T}$ and $\boldsymbol{x}(\hat{\boldsymbol{\xi}}_{1:n},\tau)$ both have some nice structure.

Let's consider a specific example of the above framework.

Example 1.3 (Decision-Aware Regularization)**.** *Suppose $c(x,\xi) = (x-\xi)^+ + (\xi - x)^+$ is the newsvendor (a.k.a. pinball loss) function.*

We let $\mathcal{T} = \mathbb{R}_+$, and for each $\tau \in \mathcal{T}$, consider fitting an exponential distribution[1] to $\hat{\boldsymbol{\xi}}_{1:n}$ with regularized maximum likelihood. Recall, the negative log-likelihood of the data is given by

$$-loglik(\lambda;\hat{\xi}_{1:n}) = -n\log(\lambda) + \lambda\sum_j \hat{\xi}_j.$$

We will use a regularization proportional to $\log(\lambda)$ *which essentially shrinks towards* 0*.*

The regularized maximum likelihood estimate is then

$$\begin{aligned}\hat{\lambda}(\tau) = \hat{\lambda}(\tau,\hat{\xi}_{1:n}) &\in \operatorname*{argmin}_{\lambda\geq 0} -n\log(\lambda) + \lambda\sum_{j=1}^{n}\hat{\xi}_j + \tau\log(\lambda)\\ &= \frac{(n-\tau)^+}{\sum_j \hat{\xi}_j}.\end{aligned}$$

Hence, we take $\hat{P}_\tau$ to be an $Exp\left(\frac{(n-\tau)^+}{\sum_j\hat{\xi}_j}\right)$ distribution.

[1]Recall $\xi \sim Exp(\lambda)$ means that $\mathbb{E}[\xi] = 1/\lambda$.

Notice then that $\boldsymbol{x}(\hat{P}_\tau(\hat{\boldsymbol{\xi}}_{1:n})) = \frac{\log 2 \sum_j \hat{\xi}_j}{n-\tau}$ *and, similarly,* $\boldsymbol{x}(\hat{\xi}_{-j}, \tau) = \frac{\log 2 \sum_{i \neq j} \hat{\boldsymbol{\xi}}_i}{n-\tau}$*. Both of these are very easy to compute, because of the way we picked our policy class.*

Since we don't know $\mathbb{P}$*, we can't evaluate* τ^{OR}*. But, we can numerically find* τ^{LOO} *by*

$$\tau^{\mathrm{LOO}} \in \operatorname*{argmin}_{0 \leq \tau \leq n} \sum_{i=1}^{n} \left[\frac{\log 2 \sum_{i \neq j} \hat{\xi}_i}{n-\tau} - \hat{\xi}_i \right]^+ + \left[\hat{\xi}_i - \frac{\log 2 \sum_{i \neq j} \hat{\xi}_i}{n-\tau} \right]^+ .$$

This problem is a simple one dimensional optimization that we can (approximately) solve by enumeration.

Remark 1.4 (Benefits of Plug-in Policies)**.** *As mentioned, plug-in policy classes are only one type of policy class. There are many others. There are at least three important benefits of plug-in policies.*

First, Eq. (1-8) *is of the same form as Problem (1-1). Thus, if one already has a specialized algorithm for solving problems of the form Problem (1-1), we can leverage the same algorithm for computing our plug-in policies. This benefit is especially important in large-scale settings.*

Second, tied to the previous point, the "structure" of the plug-in policy $\boldsymbol{x}(\mathbb{Q})$ *is often similar to the "structure" of* $\boldsymbol{x}(\mathbb{P})$ *which can be important in practice. For example, when* $\mathcal{X}$ *is a polyhedron and* $c(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{x}^\top \boldsymbol{\xi}$*, both* $\boldsymbol{x}(\mathbb{P})$ *and* $\boldsymbol{x}(\mathbb{Q})$ *will be extreme points. Other, regularization or robust optimzation approaches, might not enforce this structure. In network optimization problems, e.g., this distinction can be important, since we often need our solution to be an extreme point to be interpretable as a path or spanning tree.*

Finally, plug-in policies generalize estimate-then-optimize procedures. We often are interested in the benefits of a decision-aware approach over estimate-then-optimize. By construction, our oracle benchmark is necessarily no worse than the predict-then-optimize policy, and so might offer a way to try and quantify the benefit of decision-aware approaches. Indeed, the if $\hat{\mathbb{P}}_\tau$ *corresponds to a traditional statistical estimator for* $\mathbb{P}$*, then the gap* $\mathbb{E}\left[c(\boldsymbol{x}(\hat{\mathbb{P}}_\tau), \boldsymbol{\xi})\right] - \mathbb{E}\left[c(\boldsymbol{x}(\hat{\mathbb{P}}_\tau^{\mathrm{OR}}), \boldsymbol{\xi})\right]$ *quantifies the potential benefits of decision-aware learning for the problem structure and class* $\mathcal{T}$ *in an algorithm-independent way.*

3.4. **Other Examples.** Many, many other algorithms in data-driven optimization can be seen through the lens of proxy-objectives including:

- The SPO+ Surrogate loss [EG22]
- The Stein Correction for Small-Data, Large-Scale Linear Optimization Problems [GR21]
- The Variance Gradient Correction [GHR22]
- The Predictive to Prescriptive Method [BK20]
- Policy Learning with Doubly Robust Estimates [DLL11]
- and more.

That said, other popular algorithms are not (easily) seen as proxy-objective methods, e.g. applying stochastic gradient descent directly to Problem (1-1). These lecture notes exclusively focus on proxy-objective methods.

4. Overview of Analysis Technique

Our goal in viewing these varied algorithms through a common lens of proxy-objectives was to (hopefully) develop some common tools for their analysis. To that end, we next sketch the fundamental lemma that we will use throughout this course to prove performance guarantees for methods based on proxy-objectives.

To express the idea in a general setting, suppose we are interested in the target optimization problem

$$\boldsymbol{x}^* \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}),$$

but we instead solve the proxy objective problem

$$\hat{\boldsymbol{x}} \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \hat{f}(\boldsymbol{x}).$$

We then bound the suboptimality of $\hat{\boldsymbol{x}}$ by the approximation quality of this proxy.

Lemma 1.5 (Bounding Suboptimality by Uniform Error)**.** *We have*

$$0 \le f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \le 2 \sup_{\boldsymbol{x}\in\mathcal{X}} \left| f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}) \right|.$$

Proof. The first inequality is clear by optimality of $\boldsymbol{x}^*$. For the second, note that

$$\begin{aligned} f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^*) &= f(\hat{\boldsymbol{x}}) - \hat{f}(\hat{\boldsymbol{x}}) + \underbrace{\hat{f}(\hat{\boldsymbol{x}}) - \hat{f}(\boldsymbol{x}^*)}_{\le 0 \text{ by optimality of } \hat{\boldsymbol{x}}} + \hat{f}(\boldsymbol{x}^*) - f(\boldsymbol{x}^*) \\ &\le \left| f(\hat{\boldsymbol{x}}) - \hat{f}(\hat{\boldsymbol{x}}) \right| + \left| f(\boldsymbol{x}^*) - \hat{f}(\boldsymbol{x}^*) \right| \\ &\le 2 \sup_{\boldsymbol{x}\in\mathcal{X}} \left| f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}) \right|. \end{aligned}$$

This concludes the proof. □

Thus, to provide a performance guarantee on a proxy-objective method, it suffices to show that the proxy is a good approximation to the true function everywhere over the domain. Notice the lemma makes no assumptions on the structure of $\mathcal{X}$ or the convexity/continuity/regularity of f or $\hat{f}$.

The above lemma crucially uses the optimality of $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$. (Make sure you know where.) But it does not leverage much else about the structure of $\hat{\boldsymbol{x}}$ or $\boldsymbol{x}$, other than $\hat{\boldsymbol{x}}, \boldsymbol{x} \in \mathcal{X}$. As a consequence, in some data-driven settings, the bound can be quite bad.

Exercise: Draw a picture of $f(\cdot)$ and $\hat{f}(\cdot)$ where $\boldsymbol{x} = \hat{\boldsymbol{x}}$ but $\sup_{\boldsymbol{x}\in\mathcal{X}} \left| f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}) \right| = \infty$.

Exercise: Look back at our plug-in policy class example (cf. Example 1.3). How would we apply the lemma? What are $f(\cdot)$, $\hat{f}(\cdot)$ and $\sup_{\boldsymbol{x}\in\mathcal{X}} \left| f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}) \right|$ for that example?

References

[BK20] Dimitris Bertsimas and Nathan Kallus. "From predictive to prescriptive analytics". In: *Management Science* 66.3 (2020), pp. 1025–1044.

[DLL11] Miroslav Dudík, John Langford, and Lihong Li. "Doubly robust policy evaluation and learning". In: *arXiv preprint arXiv:1103.4601* (2011).

[EG22] Adam N Elmachtoub and Paul Grigas. "Smart "Predict, then Optimize"". In: *Management Science* 68.1 (2022), pp. 9–26.

[GHR22] Vishal Gupta, Michael Huang, and Paat Rusmevichientong. "Debiasing in-sample policy performance for small-data, large-scale optimization". In: *Operations Research* (2022). Forthcoming.

[GR21] Vishal Gupta and Paat Rusmevichientong. "Small-data, large-scale linear optimization with uncertain objectives". In: *Management Science* 67.1 (2021), pp. 220–241.

Stochastic Foundations of Decision-Aware Learning **Mini-Course 2022**

Lecture 2: Introduction to Concentration

Instructor: Vishal Gupta guptavis@usc.edu

1. Some Notation

To simplify many proofs, we will write $a \lesssim b$ to mean that there exists a universal constant C (not depending on any problem data) such that $a \leq Cb$.

So for example, when $d \geq 2$, we have that $\log(2d) \lesssim \log(d)$. (Why?) But, if we only knew that $d > 1$, we *cannot* say that $\log(2d) \lesssim \log(d)$. (Why?)

2. Review of the Chernoff Technique

Recall Markov's Inequality: For any *nonnegative* random variable X, $\mathbb{P}(X > t) \leq \mathbb{E}[X]/t$. Often we establish tail bounds on a random variable by applying Markov's inequality to a monotone transformation $f(X)$. The most common transformation is $f(t) = \exp(\theta t)$ for some (well-chosen) $\theta > 0$. This technique is usually called the Chernoff technique in textbooks. The Chernoff technique is a great general purpose approach that generally turns a bound on the mgf of a random variable $\mathbb{E}[\exp(\theta X)]$ into a tail bound.

Example 2.1 (Chernoff Bound for the Normal Distribution)**.** *Suppose $X \sim \mathcal{N}(0, \sigma^2)$. Then, for any $\theta > 0$,*

$$\begin{aligned} \mathbb{P}(X > t) &= \mathbb{P}(\exp(\theta X) > \exp(\theta t)) \\ &\leq \exp(-\theta t)\mathbb{E}[\exp(\theta X)] \\ &= \exp(-\theta t + \frac{\theta^2\sigma^2}{2}), \end{aligned}$$

where we've used the MGF of a normal distribution. We then optimize θ to obtain the best possible bound. Choosing $\theta = t/\sigma^2$ yields the bound

$$\mathbb{P}(X > t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Said differently, for any $0 < \delta < 1$, with probability at least $1 - \delta$ we have

$$X \lesssim \sigma\sqrt{\log(2/\delta)}.$$

There are tighter bounds for the tails of a normal distribution, but this bound is fairly tight.

The Chernoff technique motivates us to consider classes of random variables that admit "nice" bounds on their mgf, since we can turn such bounds into tail bounds. We next introduce a particular class of random variables with "nice" mgfs.

3. SubGaussian Random Variables

Definition 2.2. *A mean-zero random variable X is subGaussian with variance proxy σ^2 if*

$$\mathbb{E}\left[e^{\theta X}\right] \;\leq\; e^{\frac{\theta^2\sigma^2}{2}} \quad \forall \theta \in \mathbb{R}.$$

As desired, this class of random variables admit a nice tail bound:

Lemma 2.3. *If $X - \mathbb{E}[X]$ is subGaussian with variance proxy σ^2, then for any $0 < \delta < 1$, we have with probability at least $1 - \delta$,*

$$X - \mathbb{E}[X] \;\leq\; \sigma\sqrt{\log(2/\delta)}.$$

Proof. Left for the reader. (Should be immediate.) □

Example 2.4 (Normal Random Variables)**.** *If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $X - \mathbb{E}[X]$ is subGaussian with variance proxy σ^2. (Why?)*

Example 2.5 (Rademacher Random Variables)**.** *Consider a Rademacher random variable η, i.e. $\mathbb{P}(\eta = 1) = \mathbb{P}(\eta = -1) = 1/2$. On HW 1, you proved that η is subGaussian with variance proxy 1.*

Interestingly, any bounded variable is subGaussian.

Theorem 2.6 (Hoeffding's Inequality)**.** *Suppose X is a random variable such that $a \leq X \leq b$ almost surely. Then $X - \mathbb{E}[X]$ is subGaussian with variance proxy σ^2 satisfying $\sigma \lesssim (b - a)$.*

Proof. Our proof uses a technique called "symmetrization" to reduce the analysis of a complicated random to the analysis of a Rademacher variable.

Let η be a Rademacher random variable and let $\bar{X}$ be an i.i.d. copy of X. Then, for any θ,

$$\begin{aligned}\mathbb{E}[\exp(\theta(X - \mathbb{E}[X]))] &\leq \mathbb{E}\left[\exp(\theta(X - \bar{X}))\right] && \text{(Jensen's Inequality)}\\ &= \mathbb{E}\left[\exp\left(\theta\eta\left|X - \bar{X}\right|\right)\right].\end{aligned}$$

Notice crucially how the last line uses the fact that $X - \bar{X}$ is symmetric.

Now consider conditioning on $X, \bar{X}$, and apply our bound on the mgf of a Rademacher random variable to obtain,

$$\mathbb{E}[\exp(\theta(X - \mathbb{E}[X]))] \leq \mathbb{E}\left[\exp\left(\theta^2\frac{(X - \bar{X})^2}{2}\right)\right].$$

Finally, note that $a \leq X \leq b$ almost surely implies that $\left|X - \bar{X}\right| \leq (b - a)$. Hence, this last expectation is at most $\exp\left(\frac{\theta^2(b-a)^2}{2}\right)$. In other words, $X - \mathbb{E}[X]$ is subGaussian, and its variance proxy is at most $(b - a)^2$.

□

Remark 2.7. *Our proof bounded the variance proxy up to a constant. This constant is not tight. See [BLM13].*

Linear combinations of subGaussian random variables are subGaussian.

Theorem 2.8 (Linear Combinations of SubGaussian Random Variables)**.** *Suppose X_i for $i = 1, \ldots, d < \infty$ are a set of mean-zero subGaussian random variables with variance proxies σ_i^2. Then,*

(1) For any $a \in \mathbb{R}$, the r.v. aX_1 is subGaussian with variance proxy $a^2\sigma_1^2$.
(2) If the X_i are independent, then $\sum_{i=1}^d X_i$ is subGaussian with variance proxy $\sum_{i=1}^d \sigma_i^2$.

Proof. You'll prove parts ii) on your homework. Part i) is immediate from the definition. (Convince yourself of this.) □

4. Orlicz Norms

There are many equivalent definitions of subGaussian random variables. One equivalence I find useful relates subGaussian random variables to random variables with finite Ψ-Orlicz norms.

Definition 2.9 (The Ψ-Orlicz Norm)**.** *Let $\Psi(t) = \frac{1}{5}\exp(t^2)$. For any random variable X, the Ψ-Orlicz Norm of X, denoted $\|X\|_\Psi$, is given by*

$$\inf\{C > 0 : \mathbb{E}\left[\Psi(X/C)\right] \leq 1\}.$$

Importantly, $\|\cdot\|_\Psi$ is a *norm* on random variables, so it satisifes the "usual" properties of a norm (e.g., triangle-inequality). Orlicz norms can be defined with respect to other functions, but we will only focus on the case of $\Psi(\cdot)$. Some authors define $\Psi(\cdot)$ with a different constant than $\frac{1}{5}$. Note, X need not be mean-zero in this definition.

From Markov's inequality, we have that for any random variable X, with probability at least $1-\delta$,

$$|X| \lesssim \|X\|_\Psi \sqrt{\log(2/\delta)}.$$

This observation motivates the following fact that you will prove on your homework:

Theorem 2.10 (Relating SubGaussian R.V.s and the $\|\cdot\|_\Psi$.)**.** *Suppose X is a mean-zero subGaussian random variable with variance proxy σ^2. Then, $\|X\|_\Psi \lesssim \sigma$. Conversely, if X is mean-zero and $\|X\|_\Psi < \infty$, then X is subGaussian with variance proxy σ^2 such that $\sigma \lesssim \|X\|_\Psi$.*

One of the most important features of the Ψ-Orlicz Norms is that it behaves nicely under the max operation.

Theorem 2.11 (Orlicz Norm of the Max)**.** *Suppose X_i for $i = 1, \ldots, d < \infty$ are (not necessarily independent) random variables such that $\|X_i\|_\Psi < \infty$ for all i. Define $X_{\max} = \max_i X_i$. Then, for $d \geq 2$, $\|X_{\max}\|_\Psi \lesssim (\max_i \|X_i\|_\Psi)\sqrt{\log d}$.*

Proof. Fix some $C, K > 0$ to be specified later. From fundamental theorem of calculus,

$$\begin{aligned}\Psi(X_{\max}/C) &= \Psi(1) + \int_1^{\frac{X_{\max}}{C}} \Psi'(t)dt \\ &= \frac{e}{5} + \int_1^\infty \mathbb{I}\left\{t \leq \frac{X_{\max}}{C}\right\}\Psi'(t)dt\end{aligned}$$

Now note that because $\Psi(\cdot)$ is increasing,

$$t \leq \frac{X_{\max}}{C} \implies \Psi(Kt) \leq \Psi\left(\frac{KX_{\max}}{C}\right) \implies \frac{\Psi\left(\frac{KX_{\max}}{C}\right)}{\Psi(Kt)} \geq 1.$$

So, increasing the integrand, we have

$$\begin{aligned}\Psi(X_{\max}/C) &\le \frac{e}{5} + \int_1^\infty \Psi\left(\frac{KX_{\max}}{C}\right)\frac{\Psi'(t)}{\Psi(Kt)}dt \\ &\le \frac{e}{5} + \sum_{i=1}^d \int_1^\infty \Psi\left(\frac{KX_i}{C}\right)\frac{\Psi'(t)}{\Psi(Kt)}dt,\end{aligned}$$

by bounding the max by the sum. Now let $C = K\max_i \|X_i\|_\Psi$ and take expectations of both sides. We can pass the expectation inside the integral by the montone convergence theorem. We thus have,

$$\begin{aligned}\mathbb{E}\left[\Psi(X_{\max}/C)\right] &\le \frac{e}{5} + d\int_1^\infty \frac{\Psi'(t)}{\Psi(Kt)} \\ &= \frac{e}{5} + 2d\int_1^\infty te^{-(K^2-1)t^2}dt \\ &= \frac{e}{5} + \frac{d}{K^2-1}e^{-(K^2-1)}.\end{aligned}$$

Taking $K = \sqrt{2+\log d}$ makes this last expression at most 1. This shows $\|X_{\max}\|_\Psi \le \sqrt{2+\log d} \sim \sqrt{\log d}$. □

Next week, we will show how to use the above tools to bound prove a performance guarantee for SAA. For now, let's see a quick toy example.

Example 2.12. *Suppose $X_1 \sim \mathcal{N}(10, \sigma^2)$ and $X_2 \sim$ Unif$[-3\sigma, 3\sigma]$, not necessarily independent with $\sigma \ge 1$. Define $Y = |X_1| + 2X_2$. We will prove a tail bound for Y.*

Our strategy will be to bound $\|Y\|_\Psi$ and then use Theorem 2.10. Then, by triangle inequality,

$$\|Y\|_\Psi \;\le\; \|\,|X_1|\,\|_\Psi + 2\|X_2\|_\Psi.$$

By Hoeffding's inequality, X_2 is subGaussian with variance proxy $\lesssim \sigma^2$, so Theorem 2.10 shows $\|X_2\|_\Psi \lesssim \sigma$.

Similarly, if we let $Z \sim \mathcal{N}(0, \sigma^2)$, then $|X_1| \le 10 + |Z|$. By Theorem 2.10 $\|\,|Z|\,\|_\Psi = \|Z\|_\Psi \lesssim \sigma$. Hence $\|\,|X|\,\|_\Psi \lesssim 10 + \sigma \lesssim \sigma$.

Combining shows $\|Y\|_\Psi \sim \sigma$, so that with proability at least $1-\delta$,

$$|Y| \lesssim \sigma\sqrt{\log(1/\delta)}.$$

5. (Time Permitting): McDiarmid's Inequality

McDiarmid's Inequality is one of the most fundamental results in concentration. It is one way to formalize an important intuition in concentration: a function of many independent random variables that doesn't depend too strongly on any one variable will concentrate at its expectation.

Before stating McDiarmid's inequality, we need to define a function with bounded differences.

Definition 2.13 (Bounded Differences)**.** *Suppose $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is a function of n random variables where $\boldsymbol{x}_i \in \mathcal{X}_i$ for each i. We say $f(\cdot)$ satisfies the bounded differences condition with respect to*

constants $c_1, \dots, c_n$ *if for each* $i = 1, \dots, n$

$$\sup_{\bar{x} \in \mathcal{X}_i} |f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) - f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{i-1}, \bar{x}, \boldsymbol{x}_{i+1}, \dots, n)| \le c_i.$$

Theorem 2.14 (McDiarmid's Inequality)**.** *Suppose* $X_i \in \mathcal{X}_i$ *for* $i = 1, \dots, n$ *are independent random variables (not necessarily identically distributed) and consider a function* $f(X_1, \dots X_n)$*. Suppose* $f(\cdot)$ *satisfies the bounded difference inequality with respect to a constants* c_i *for* $i = 1, \dots, n$*. Then,* $f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) - \mathbb{E}\left[f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)\right]$ *is subGaussian with variance proxy* $\lesssim \|\boldsymbol{c}\|_2^2$*.*

In particular, for any $0 < \delta < 1$*, with probability at least* $1 - \delta$

$$|f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) - \mathbb{E}\left[f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)\right]| \lesssim \|\boldsymbol{c}\|_2 \sqrt{\log(1/\delta)}.$$

Proof. We provide an elementary proof based on the Doob martingale. Define

$$Z_i = \mathbb{E}\left[f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) \mid \boldsymbol{X}_{1:i}\right] - \mathbb{E}\left[f(\boldsymbol{X}_1, \dots, X_n) \mid \boldsymbol{X}_{1:(i-1)}\right].$$

Then, $f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) - \mathbb{E}\left[f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)\right] = \sum_{i=1}^n Z_i$. (Check this!). Moreover, Z_i is only a function of $X_{1:i}$.

Now let $\bar{\boldsymbol{X}}_i$ be an i.i.d copy of $\boldsymbol{X}_i$. Then, because of the independence,

$$Z_i = \mathbb{E}\left[f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) - f(\boldsymbol{X}_1, \dots, \boldsymbol{X}_{i-1}, \bar{\boldsymbol{X}}_i, \boldsymbol{X}_{i+1}, \dots, \boldsymbol{X}_n) \mid \boldsymbol{X}_{1:i}\right],$$

and by the bounded differences condition, $|Z_i| \le c_i$.

By the tower rule,

$$\mathbb{E}\left[\exp(\theta \sum_{i=1}^n Z_i)\right] = \mathbb{E}\left[\exp(\theta \sum_{i=1}^{n-1} Z_i)\mathbb{E}\left[\exp(\theta Z_n) \mid \boldsymbol{X}_{1:(n-1)}\right]\right].$$

We can then apply Hoeffding's inequality (conditionally) to conclude there exists a universal constant C such that

$$\mathbb{E}\left[\exp(\theta \sum_{i=1}^{n-1} Z_i)\mathbb{E}\left[\exp(\theta Z_n) \mid \boldsymbol{X}_{1:(n-1)}\right]\right] \;\le\; e^{C\theta^2 c_i^2}\mathbb{E}\left[\exp(\theta \sum_{i=1}^{n-1} Z_i)\right].$$

We apply this process repeatedly and conclude that

$$\mathbb{E}\left[\exp(\theta(f(\boldsymbol{X}_{1:n}) - \mathbb{E}\left[f(\boldsymbol{X}_{1:n})\right]))\right] \;\le\; \exp(C\theta^2\|\boldsymbol{c}\|_2^2).$$

This completes the proof. □

6. Some Remarks

The theory of subGaussian random variables is incredibly rich. I highly encourage you to read the "Equivalent Characterizations of SubGaussian Random Variables' in [Wai19]. While I don't often use this theorem directly, it's very good for developing intuition about when something might subGaussian.

Orlicz norms are a handy tool (in my opinion) for manipulating potentially non-independent random variables using reasoning based on norms. While we focus on the $\|\cdot\|_\Psi$, Orlicz norms for other functions are often necessary when the tails don't have subGaussian behavior. Other common

Orlicz Norms used in the literature correspond to the functions $\Psi(t) = \exp(t)$ or $\Psi(t) = t^p$ for some $p \geq 1$. We won't use these other norms in this class.

Not all random variables are subGaussian. A much richer class of random variables are those with finite moment generating function, and such random variables are described subExponential or subGamma. [BLM13] treats subGamma random variables fairly well, and shows why you sometimes need to consider them. Essentially, if you start ending up with tail bounds that depend on $\log(1/\delta)$ instead of $\sqrt{1/\log(\delta)}$, you probably need to work with subGamma random variables.

There are *many* generalizations of McDiarmid's inequality. [BLM13] has some good ones, but there are still new ones being published in modern conferences (e.g. [MP21]). One of the ones I've found most useful in my research is [Com15]. The proof is particularly simple and beautiful (and short!). [Wai19] gives some excellent examples of how to use McDiarmid's inequality.

References

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[Com15] Richard Combes. "An extension of McDiarmid's inequality". In: *arXiv preprint arXiv:1511.05240* (2015).

[MP21] Andreas Maurer and Massimiliano Pontil. "Concentration inequalities under sub-Gaussian and sub-exponential conditions". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 7588–7597.

[Wai19] Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Vol. 48. Cambridge University Press, 2019.

Stochastic Foundations of Decision-Aware Learning **Mini-Course 2022**

Lecture 3: Towards Performance Bounds: Packing

Instructor: Vishal Gupta guptavis@usc.edu

1. Performance Bound: SAA for Finite Feasible Regions

Using the tools we've developed so far, we prove a performance bound on SAA. Recall, our target optimization is

$$\boldsymbol{x}^* \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}\left[c(\boldsymbol{x},\boldsymbol{\xi})\right]$$

where $\boldsymbol{\xi} \sim \mathbb{P}$, and $\mathbb{P}$ is unknown. We have access to data $\hat{\boldsymbol{\xi}}_j \sim \mathbb{P}$ i.i.d. for $j = 1, \ldots, n$. We consider the SAA policy

$$\hat{\boldsymbol{x}}_n \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\sum_{j=1}^{n} c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j).$$

We prove the following theorem from our first lecture.

Theorem 3.1 (SAA for Finite Feasible Regions)**.** *Suppose $|\mathcal{X}| < \infty$ and that there exists $c_{\max} < \infty$ such that $\max_{\boldsymbol{x}\in\mathcal{X}} |c(\boldsymbol{x},\boldsymbol{\xi})| \le c_{\max}$ almost surely. Then, for any $0 < \delta < 1$, with probability at least $1-\delta$,*

$$0 \;\le\; \mathbb{E}\left[c(\hat{\boldsymbol{x}}_n,\boldsymbol{\xi}) \mid \hat{\boldsymbol{\xi}}_{1:n}\right] - \mathbb{E}\left[c(\boldsymbol{x}^*,\boldsymbol{\xi})\right] \;\lesssim\; c_{\max}\sqrt{\frac{\log(|\mathcal{X}|)\log(2/\delta)}{n}}.$$

Proof of Theorem 3.1. By Lemma "Bounding Suboptimality by Uniform Error" from Lecture 1, it suffices to bound

$$\max_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{1}{n}\sum_{j=1}^{n} c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\boldsymbol{\xi})\right]\right| \;=\; \max_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{1}{n}\sum_{j=1}^{n}\left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right]\right)\right|.$$

Fix some $\boldsymbol{x} \in \mathcal{X}$. Note that for each j, $\left|c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right| \lesssim c_{\max}$. Hence, by Hoeffding's inequality, $c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right]$ is subGaussian with variance proxy $\lesssim c_{\max}^2$. Since the j are independent, it follows that $\frac{1}{n}\sum_{j=1}^{n} c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right]$ is subGaussian with variance proxy $\lesssim c_{\max}^2/n$. (Why?)

By the relation between subGaussian variables and the Ψ-Orlicz norm,

$$\left\|\frac{1}{n}\sum_{j=1}^{n} c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right]\right\|_{\Psi} \lesssim c_{\max}/\sqrt{n}.$$

From the maxima of Orlicz Norms, it follows that

$$\left\| \max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{1}{n}\sum_{j=1}^{n} \left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) \right] \right) \right| \right\|_{\Psi} \lesssim c_{\max}\sqrt{\frac{\log(|\mathcal{X}|)}{n}}.$$

The result then follows from the tail bound of a r.v. with finite Ψ-norm. □

It's worth stressing how short and simple the above proof is. You should compare it to the proof in [KSH02] which establishes (qualitatively) similar results.

1.1. **An Alternate Proof Based on Symmetrization.** Before proceeding, we present a different proof of the above theorem. This alternate version more closely mirrors traditional proofs in the literature and paves the way for more advanced results in the course.

Before establishing the proof, we establish the following auxiliary lemma that is of independent interest.

Lemma 3.2 (Symmetrization Lemma)**.** *For any convex, increasing function $\Phi(\cdot)$,*

$$\mathbb{E}\left[\Phi\left(\max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{1}{n}\sum_{j=1}^{n} \left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) \right] \right) \right| \right) \right] \;\le\; \mathbb{E}\left[\Phi\left(2\max_{\boldsymbol{c}\in\mathcal{F}} \left| \frac{1}{n}\sum_{j=1}^{n} \eta_j c_j \right| \right) \right], \tag{3-11}$$

where η_j for $j = 1,\dots,n$ are independent Rademacher random variables and

$$\mathcal{F} = \mathcal{F}(\hat{\boldsymbol{\xi}}_{1:n}) = \{(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_1),\dots,c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_n))^\top : \boldsymbol{x}\in\mathcal{X}\} \subseteq \mathbb{R}^n.$$

Proof of Lemma. Let $\bar{\boldsymbol{\xi}}_j$ for $j = 1,\dots,n$ be a second i.i.d. sample from $\mathbb{P}$ and fix the function $\Phi(\cdot)$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\Phi\left(\max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{1}{n}\sum_{j=1}^{n} \left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) \right] \right) \right| \right) \right] \\
&\quad= \mathbb{E}\left[\Phi\left(\max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{1}{n}\sum_{j=1}^{n} \left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\bar{\boldsymbol{\xi}}_j) \mid \hat{\boldsymbol{\xi}}_{1:n} \right] \right) \right| \right) \right] \\
&\quad\le \mathbb{E}\left[\Phi\left(\max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{1}{n}\sum_{j=1}^{n} \left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - c(\boldsymbol{x},\bar{\boldsymbol{\xi}}_j) \right) \right| \right) \right] \qquad \text{(Jensen's Inequality).}
\end{aligned}$$

$$\mathbb{E}\left[\Phi\left(\max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{1}{n}\sum_{j=1}^{n} \left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - c(\boldsymbol{x},\bar{\boldsymbol{\xi}}_j) \right) \right| \right) \right] \;\le\; \mathbb{E}\left[\Phi\left(\max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{1}{n}\sum_{j=1}^{n} \eta_j \left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - c(\boldsymbol{x},\bar{\boldsymbol{\xi}}_j) \right) \right| \right) \right]$$

Now, focus on the argument of $\Phi(\cdot)$. Observe

$$\begin{aligned}\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j\left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)-c(\boldsymbol{x},\bar{\boldsymbol{\xi}}_j)\right)\right| &\le \max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j\zeta(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right|+\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j\zeta(\boldsymbol{x},\bar{\boldsymbol{\xi}}_j)\right| \\ &= \frac{1}{2}\left(2\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j\zeta(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right|+2\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j\zeta(\boldsymbol{x},\bar{\boldsymbol{\xi}}_j)\right|\right)\end{aligned}$$

Hence, increasing the argument of $\Phi(\cdot)$ and using convexity shows that

$$\begin{aligned}&\mathbb{E}\left[\Phi\left(\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j\left(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)-c(\boldsymbol{x},\bar{\boldsymbol{\xi}}_j)\right)\right|\right)\right] \\ &\quad\le \frac{1}{2}\mathbb{E}\left[\Phi\left(2\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right|\right)\right]+\frac{1}{2}\mathbb{E}\left[\Phi\left(2\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j c(\boldsymbol{x},\bar{\boldsymbol{\xi}}_j)\right|\right)\right] \\ &\quad= \mathbb{E}\left[\Phi\left(2\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right|\right)\right].\end{aligned}$$

This completes the proof. □

Lemma 3.2 is used regularly in the literature when analyzing sums of independent random variables. You should study the set $\mathcal{F}$ defined in the lemma. This set is random (it depends on the data) and also depends on the particular cost function and feasible region.

IMPORTANT: *Our lemma did not require that $|\mathcal{X}|<\infty$!*

Equipped with the Lemma 3.2, we provide our alternate proof of Theorem 3.1.

Alternate Proof of Theorem 3.1. We again focus on bounding the Ψ Orlicz norm of

$$\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^{n}c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)-\mathbb{E}\left[c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right]\right|.$$

Motivated by Eq. (3-11), we'll first study

$$\max_{\boldsymbol{c}\in\mathcal{F}(\hat{\boldsymbol{\xi}}_{1:n})}\left|\frac{1}{n}\boldsymbol{c}^\top\boldsymbol{\eta}\right|.$$

There are two sources of randomness here: $\hat{\boldsymbol{\xi}}_{1:n}$ and the Rademacher variables $\eta_{1:n}$. We first argue conditionally, conditioning on $\hat{\boldsymbol{\xi}}_{1:n}$.

Specifically, for any fixed $\boldsymbol{c}$, $\frac{1}{n}\boldsymbol{\eta}^\top\boldsymbol{c}$ is subGaussian and $\|\frac{1}{n}\boldsymbol{\eta}^\top\boldsymbol{c}\|_\Psi \lesssim \frac{1}{n}\|\boldsymbol{c}\|_2$. (Why?)

Hence, from our result on the Orlicz norm of the max, it follows that the *conditional* Orlicz norm satisfies

$$\left\|\max_{\boldsymbol{c}\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j c_j\right|\right\|_{\Psi|\hat{\boldsymbol{\xi}}_{1:n}} \lesssim \frac{\sqrt{\log|\mathcal{F}|}}{n}\cdot\max_{\boldsymbol{c}\in\mathcal{F}}\|\boldsymbol{c}\|_2 \lesssim c_{\max}\sqrt{\frac{\log|\mathcal{F}|}{n}}.$$

Said differently, there exists a universal constant K such that

$$\left\| \max_{\boldsymbol{c} \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^{n} \eta_j c_j \right| \right\|_{\Psi | \hat{\boldsymbol{\xi}}_{1:n}} = K c_{\max} \sqrt{\frac{\log |\mathcal{F}|}{n}}.$$

For convenience, let $J = J(\hat{\boldsymbol{\xi}}_{1:n}) = K c_{\max} \sqrt{\frac{\log |\mathcal{F}|}{n}}$. Then, taking an expectation of the definition of the Orlicz Norm, we've thus far shown

$$\mathbb{E}\left[\exp\left(\frac{\max_{\boldsymbol{c} \in \mathcal{F}} \left(\frac{1}{n} \sum_{j=1}^{n} \eta_j c_j \right)^2}{J(\hat{\boldsymbol{\xi}}_{1:n})^2} \right) \right] \leq 5. \tag{3-12}$$

It would be great at this point if we could apply Lemma 3.2 with the convex function $t \mapsto \exp\left(\frac{t^2}{2J^2}\right)$ and then use Eq. (3-12) to finish the proof. Unfortunately, we can't do this because $J = J(\hat{\boldsymbol{\xi}}_{1:n})$ is random. However, since

$$|\mathcal{F}| \leq |\mathcal{X}|, \tag{3-13}$$

J is upper bounded by $K c_{\max} \sqrt{\frac{\log |\mathcal{X}|}{n}}$. Hence, we can instead take

$$\Phi(t) \equiv \exp\left(\frac{t^2}{2K^2 c_{\max}^2 \frac{\log |\mathcal{X}|}{n}} \right).$$

Then, applying Lemma 3.2 with this convex, increasing function, and using Eq. (3-12) shows

$$\mathbb{E}\left[\Phi\left(\max_{\boldsymbol{x} \in \mathcal{X}} \left| \frac{1}{n} \sum_{j=1}^{n} \left(c(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_j) \right] \right) \right| \right) \right] \leq 5.$$

Applying Markov's Inequality completes the proof. □

The above proof might look a lot more complicated, but really the only "hard" part is establishing Lemma 3.2.

Philosophically, the advantage of the above proof is we replaced the analysis of a (potentially complicated) random variable $c(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_j)$ with analysis of i) Rademacher random variables η_j and ii) the set $\mathcal{F}(\hat{\boldsymbol{\xi}}_{1:n})$. The Rademacher random variables are simple to analyze, and in most arguments of this form, understanding $\mathcal{F}(\hat{\boldsymbol{\xi}}_{1:n})$ comes down to geometry instead of probability.

Practically, the above proof really relies on the size of the set $\mathcal{F}$, not the size of the set $\mathcal{X}$. In some cases, it's easier to analyze one set rather than the other. On your homework, you'll see an example where $|\mathcal{X}| = \infty$, but $|\mathcal{F}| < \infty$, so in fact you *have* to analyze things with the second proof.

The downside of the alternate proof is that Eq. (3-11) *heavily* leverages the fact that we're looking at a sum of independent random variables. By contrast, our first proof just required us to manipulate tail bounds and might be adapted more easily to other settings.

2. Beyond Finite Feasible Regions: Packing in Metric Spaces

A lot of the theory of uniform laws of large numbers involves generalizing the machinery of the above proofs to the case where $\mathcal{X}$ (or $\mathcal{F}$) is not finite by approximating it by a finite subset. To this end, we introduce the notion of a packing and covering in Euclidean spaces.

Definition 3.3 (Packing Number)**.** *Given a set $\mathcal{F} \subseteq \mathbb{R}^n$ and a norm $\|\cdot\|$ on $\mathbb{R}^n$, the ϵ packing number of $\mathcal{F}$ (denoted $D(\epsilon, \mathcal{F}, \|\cdot\|)$) is the maximum number of points in $\mathcal{F}$ such that every pair of points is at least ϵ far apart with respect to $\|\cdot\|$.*

Definition 3.4 (Covering Number)**.** *Given a set $\mathcal{F} \subseteq \mathbb{R}^n$ and a norm $\|\cdot\|$ on $\mathbb{R}^n$, the ϵ-covering number of $\mathcal{F}$ (denoted $N(\epsilon, \mathcal{F}, \|\cdot\|)$) is the smallest number of closed balls of radius ϵ whose union contains $\mathcal{F}$.*

Exercise: Draw a picture in your notes to understand the definitions.

It's possible to generalizes both definitions to general metric spaces, but we won't need it. In this course, we'll mostly be concerned with packings, but the two nubmers are closely related. Indeed, you should prove to yourself that for any set $\mathcal{F}$,

$$N(\epsilon, \mathcal{F}, \|\cdot\|) \;\leq\; D(\epsilon, \mathcal{F}, \|\cdot\|) \;\leq\; N(\epsilon/2, \mathcal{F}, \|\cdot\|).$$

Example 3.5 (Packing an Interval)**.** *What is the ϵ packing number of $[0,1]$ with respect to ℓ_2 norm? Are you sure?*

Exercise: Let $\mathcal{F}$ be a compact subset of $\mathbb{R}^n$. Draw a plot where the x-axis is ϵ and the y-axis is $D(\epsilon, \mathcal{F}, \|\cdot\|_2)$.

In our proofs, we will mostly use the ℓ_2 packing number. You'll prove some useful facts about packing numbers on your homework. For now, one very important fact is the following:

Theorem 3.6 (Packing of Euclidean Balls)**.** *The ϵ packing number of a ball of radius $R > \epsilon$ in $\mathbb{R}^d$ with respect to ℓ_2 satisfies*

$$D(\epsilon, B(\mathbf{0}, R), \|\cdot\|_2) \;\leq \left(\frac{3R}{\epsilon}\right)^d.$$

Proof. Consider some packing of $B(\mathbf{0}, R)$ and denote the points as $f_1, \ldots, f_m$ where $m = D(\epsilon, B(\mathbf{0}, R), \|\cdot\|_2\|)$. Now consider balls of radius $\epsilon/2$ centered at each f_i. By construction, these balls are disjoint. Their total volume is $Km2^{-d}\epsilon^d$ for some constant K.

Now each f_i lies within a distance R of f_1 because the f_i are in $B(\mathbf{0}, R)$. Hence, the ball $B(f_i, \epsilon/2)$ lies within the larger ball $B(f_1, 3R/2)$ for each i. This larger ball has volume $KR^d 1.5^d$. Comparing completes the proof. □

In class, we'll leverage packing numbers to extend our proof of SAA to non-finite feasible regions with small packing numbers.

References

[KSH02] Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de-Mello. "The sample average approximation method for stochastic discrete optimization". In: *SIAM Journal on Optimization* 12.2 (2002), pp. 479–502.

Stochastic Foundations of Decision-Aware Learning **Mini-Course 2022**

Lecture 4: More Uniform Laws of Large Numbers

Instructor: Vishal Gupta guptavis@usc.edu

1. Review

Let's recap: So far, we've shown how analyzing a data-driven method based on a proxy-objective reduces to studying the uniform approximation error between the true objective and proxy-objective. In the case of SAA, this amounted to studying the uniform error between a sample average and a true expectation over the feasible region.

In this lecture, we're going to push this idea further to develop stronger bounds on this uniform error. These will translate into other performance guarantees for SAA.

2. SAA with Bounded Feasible Regions and Lipschitz Cost Functions

As a first example, we'll study our usual stochastic optimization

$$\boldsymbol{x}^* \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}\left[c(\boldsymbol{x},\boldsymbol{\xi})\right]$$

under the assumptions that $\boldsymbol{x} \mapsto c(\boldsymbol{x},\boldsymbol{\xi})$ is L-Lipschitz almost surely, and $\mathcal{X} \subseteq \mathbb{R}^d$ is a bounded feasible region.

Definition 4.1. *A set $\mathcal{X} \subseteq \mathbb{R}^d$ is bounded if $R \equiv \frac{1}{2}\sup_{\boldsymbol{x},\bar{\boldsymbol{x}}\in\mathcal{X}} \|\boldsymbol{x}-\bar{\boldsymbol{x}}\| < \infty$. In that case, we say R is* radius *of $\mathcal{X}$.*

Recall from our earlier discussion on packing, $\log D(\epsilon,\mathcal{X}) \le d\log(3R/\epsilon)$. Finally, let

$$\hat{\boldsymbol{x}}_n \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\sum_{j=1}^n c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)$$

be the SAA solution.

We then have the following theorem:

Theorem 4.2 (SAA for Bounded Feasible Regions, Lipschitz Cost Functions)**.** *Suppose $\mathcal{X}$ is bounded with radius $R < \infty$, and $\boldsymbol{x} \mapsto c(\boldsymbol{x},\boldsymbol{\xi})$ is L-Lipschitz almost surely. Further, assume $\max_{\boldsymbol{x}\in\mathcal{X}} |c(\boldsymbol{x},\boldsymbol{\xi})| \le c_{\max}$ almost surely and $n \ge 2$.*

Then,

$$\left\| \max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{1}{n}\sum_{j=1}^n c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right] \right| \right\|_\Psi \lesssim (L + c_{\max}\sqrt{\log(3R+2)})\sqrt{\frac{d\log n}{n}}.$$

Consequently, for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$0 \le \mathbb{E}\left[c(\hat{\boldsymbol{x}}_n, \boldsymbol{\xi}) \mid \hat{\boldsymbol{\xi}}_{1:n}\right] - \mathbb{E}\left[c(\boldsymbol{x}^*, \boldsymbol{\xi})\right] \le (L + c_{\max}\sqrt{\log(3R+2)})\sqrt{\frac{d\log n}{n}} \cdot \sqrt{\log(2/\delta)}.$$

Proof. By the Symmetrization Lemma, it will suffice to bound

$$\frac{1}{n}\left\|\max_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{c}(\boldsymbol{x})^\top \boldsymbol{\eta}\right\|_\Psi$$

where $\boldsymbol{c}(\boldsymbol{x}) \equiv (c(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_1), \ldots, c(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_n))^\top \in \mathbb{R}^n$, and $\boldsymbol{\eta} \in \mathbb{R}^n$ is a vector of i.i.d. Rademacher random variables. (Convince yourself this is true.)

Now fix some $0 < \epsilon < 1$ (to be specified later), and let $\mathcal{X}_0 = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ be an ϵ-packing of $\mathcal{X}$. By our earlier remarks,

$$\log(m) \le d\log(3R/\epsilon) \lesssim d\log(e+R)\log(2/\epsilon).$$

Then, for any $\boldsymbol{x} \in \mathcal{X}$, there exists a $\boldsymbol{x}_k \in \mathcal{X}_0$ such that $\|\boldsymbol{x} - \boldsymbol{x}_k\| \le \epsilon$. (Why?) For such a k, we have that

$$\begin{aligned}
\|\boldsymbol{c}(\boldsymbol{x}) - \boldsymbol{c}(\boldsymbol{x}_k)\|^2 &= \sum_{j=1}^n (c(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_j) - c(\boldsymbol{x}_k, \hat{\boldsymbol{\xi}}_j))^2 \\
&\le \sum_{j=1}^n L^2 \|\boldsymbol{x} - \boldsymbol{x}_k\|^2 \\
&\le L^2\epsilon^2 n.
\end{aligned}$$

This implies $\|\boldsymbol{c}(\boldsymbol{x}) - \boldsymbol{c}(\boldsymbol{x}_k)\|_2 \le L\epsilon\sqrt{n}$.

Now write,

$$\begin{aligned}
\boldsymbol{c}(\boldsymbol{x})^\top \boldsymbol{\eta} &= (\boldsymbol{c}(\boldsymbol{x}) - \boldsymbol{c}(\boldsymbol{x}_k))^\top \boldsymbol{\eta} + \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{\eta} \\
&\le L\epsilon\sqrt{n}\|\boldsymbol{\eta}\| + \boldsymbol{c}(\boldsymbol{x}_k)^\top \boldsymbol{\eta} && \text{(Cauchy-Schwarz)} \\
&\le L\epsilon\sqrt{n}\|\boldsymbol{\eta}\| + \max_{i=1,\ldots,m} \boldsymbol{c}(\boldsymbol{x}_i)^\top \boldsymbol{\eta}.
\end{aligned}$$

Notice, $\|\boldsymbol{\eta}\|_2 = \sqrt{n}$ almost surely. Hence, if we take the maximum over $\boldsymbol{x} \in \mathcal{X}$ of both sides we have that

$$\max_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{c}(\boldsymbol{x})^\top \boldsymbol{\eta} \;\le\; L\epsilon n + \max_{i=1,\ldots,m} \boldsymbol{c}(\boldsymbol{x}_i)^\top \boldsymbol{\eta}.$$

Now take the conditional Ψ Orlicz norm (conditional on $\hat{\boldsymbol{\xi}}_{1:n}$) of both sides. For each i, $\|\boldsymbol{c}(\boldsymbol{x}_i)^\top \boldsymbol{\eta}\|_{\Psi|\hat{\boldsymbol{\xi}}_{1:n}} \lesssim \|\boldsymbol{c}(\boldsymbol{x}_i)\|_2 \lesssim c_{\max}\sqrt{n}$. Hence, using our result for the Orlicz Norm of the max, we have that

$$\left\|\max_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{c}(\boldsymbol{x})^\top \boldsymbol{\eta}\right\|_{\Psi|\hat{\boldsymbol{\xi}}_{1:n}} \;\lesssim\; L\epsilon n + c_{\max}\sqrt{n\log m} \;\lesssim\; L\epsilon n + c_{\max}\sqrt{nd\log(3R)\log(2/\epsilon)}.$$

We are now free to choose $\epsilon > 0$ to optimize the bound. Optimizing it exactly seems tricky, so instead we pick a (suboptimal) choice of $\epsilon = \frac{1}{\sqrt{n}}$. This yields,

$$\begin{aligned}\left\|\max_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{c}(\boldsymbol{x})^\top \boldsymbol{\eta}\right\|_{\Psi|\hat{\boldsymbol{\xi}}_{1:n}} &\lesssim L\sqrt{n} + c_{\max}\sqrt{nd\log(e+R)\log(2\sqrt{n})} \\ &\lesssim (L + c_{\max}\sqrt{\log(e+R)})\sqrt{\frac{d\log n}{n}}.\end{aligned}$$

By writing out the definition of the conditional Ψ-norm above and taking an expectation with respect to $\hat{\boldsymbol{\xi}}_{1:n}$, we find the unconditional Ψ-norm above is also bounded by the same quantity.

Applying the Symmetrization Lemma then shows

$$\left\|\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{n}\sum_{j=1}^n c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_j)\right|\right\|_\Psi \;\leq\; (L + c_{\max}\sqrt{\log(e+R)})\sqrt{\frac{d\log n}{n}}.$$

This proves the first statement. The final statement follows from Markov's Inequality. □

Some remarks: It's worth noting that we can view the approximating finite set $\mathcal{X}_0$ either as a packing of $\mathcal{X}$, or we can view the induced set

$$\mathcal{F}_0 = \{\Big(c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_1),\ldots,c(\boldsymbol{x},\hat{\boldsymbol{\xi}}_n)\Big)^\top : \boldsymbol{x}\in\mathcal{X}_0\}$$

as a $L\epsilon\sqrt{n}$-covering of $\mathcal{F}(\hat{\boldsymbol{\xi}}_{1:n})$. This second perspective is perhaps more illuminating for the results below.

Secondly, the assumptions on $\mathcal{X}$ weren't strictly necessary. What we really needed was that the packing numbers of $\mathcal{X}$ didn't grow too rapidly as $\epsilon \to 0$. In the above case, they looked roughly like ϵ^{-d}, and this was sufficient so long as $d \ll n$. We could make other assumptions to induce "small" packing numbers for $\mathcal{X}$, or, equivalently, small packing numbers for $\mathcal{F}$.

3. More Advanced Results

As is hopefully clear by now, many performance guarantees come down to bounding the maximal deviation of a quantity like

$$Z = Z(\hat{\boldsymbol{\xi}}_{1:n}) = \sup_{t\in\mathcal{T}}\left|\sum_{j=1}^n f_j(t,\hat{\boldsymbol{\xi}}_j) - \mathbb{E}\left[f_j(t,\hat{\boldsymbol{\xi}}_j)\right]\right| \tag{4-14}$$

for some functions f_j and independent random variables $\hat{\boldsymbol{\xi}}_j$. Below we quote some stronger results from [Pol90] that bound the tails of Z when the $\hat{\boldsymbol{\xi}}_j$ are independent, but not necessarily identically distributed. We won't prove these results – they rely on some ideas we haven't talked about like chaining – but they're useful to know. When you write a paper, you should use these results from [Pol90].

To state the results, first, recall that symmetrization encouraged us to think about the (random) set

$$\mathcal{F} = \mathcal{F}(\hat{\boldsymbol{\xi}}_{1:n}) = \left\{\Big(f_1(t,\hat{\boldsymbol{\xi}}_1),\ldots,f_n(t,\hat{\boldsymbol{\xi}}_n)\Big)^\top : t\in\mathcal{T}\right\}.$$

We define the entropy integral of $\mathcal{F}$ by

$$J = J(\hat{\boldsymbol{\xi}}_{1:n}) \equiv 9 \int_0^{\delta} \sqrt{\log D(\epsilon, \mathcal{F})} d\epsilon \quad \text{where } \delta = \delta(\hat{\boldsymbol{\xi}}_{1:n}) = \sup_{\boldsymbol{f} \in \mathcal{F}} \|\boldsymbol{f}\|_2.$$

On first reading, entropy integrals are intimidating. Remember, though, that this integral is not defined probabilistically. For a fixed $\hat{\boldsymbol{\xi}}_{1:n}$, it's just an ordinary integral from calculus. Intuitively, packing numbers measure the complexity of $\mathcal{F}$ at scale ϵ. The entropy integral measures complexity at varying scales, and, if the packing numbers "blow up" too fast as $\epsilon \to 0$, then this integral diverges.

Using a standard change of variables from calculus,

$$J = 9\delta \int_0^1 \sqrt{\log D(\delta\epsilon, \mathcal{F})} d\epsilon.$$

Perhaps surprisingly, although this integral depends on the random set $\mathcal{F}$ and random "radius" δ, we can often upper bound it by a deterministic constant, so the only randomness in J comes from the leading δ term.

We give one such example below:

Lemma 4.3 (Entropy Integral for Euclidean Classes)**.** *Suppose that $\mathcal{F}$ is* Euclidean*, i.e., there exists constants A and W such that with probability* 1*,*

$$D(\epsilon\delta, \mathcal{F}) \;\leq\; A\epsilon^{-W} \quad 0 < \epsilon < 1.$$

Then,

$$J \;\lesssim\; \delta \cdot \frac{\log A + W/2}{\sqrt{\log A}}.$$

Proof. You'll prove this on your homework. (It's just an exercise in evaluating/bounding an integral!) If you're really struggling, look at Theorem A.2 in [GR21], but try your best before looking it up. □

Regardless of whether $\mathcal{F}$ is Euclidean, we get nice bounds on Z in terms of bounds on J.

Theorem 4.4 (Bounds on Suprema of Empirical Process)**.** *Fix any $0 < \beta < 1$.*

(1) (When J bounded by a constant) Suppose there exists a deterministic constant K such that $J \leq K$ almost surely. Then, with probability at least $1 - \beta$,

$$Z \;\lesssim\; K\sqrt{log(2/\beta)}.$$

(2) (When J is SubGaussian) Suppose that $\|J\|_{\Psi} < \infty$. Then, with probability at least $1 - \beta$,

$$Z \;\lesssim\; \|J\|_{\Psi} \log(2/\beta).$$

(3) (When J has finite p^{th} moment) Suppose that $\|J\|_{\mathcal{L}^p} \equiv \mathbb{E}\,[J^p]^{1/p} < \infty$. Then, with probability at least $1 - \beta$,

$$Z \;\lesssim\; \sqrt{p}\beta^{-1/p}\|J\|_{\mathcal{L}^p}.$$

Of course, when $\mathcal{F}$ is Euclidean we can simplify the results above appropriately. In particular, in that case the bounds above really depend on the behavior of δ, i.e, how big do the largest elements of the process get? I'll leave the details to you...

Let's step back: What have we done? The above theorem is a general purpose tool. It's important to us because we can often reduce the analysis of a performance guarantee for an optimization method to studying the tails of a random variable like Z. and then we can use the above theorem to get a bound.

You'll follow this recipe on your homework to reprove a stronger performance guarantee under the setting of Theorem 4.2. If you do it right, you'll find the result is only really stronger by a logarithmic factor.... And this insight is important (to me). Namely, if you use some of the strongest methods available in empirical process, you often only get small improvements over the more naive (by-hand) approaches. And it's often more straightforward to adapt these "by-hand" approaches when your problem doesn't *exactly* match the setting of the theorems in a textbook. So, don't be afraid to do things by hand.

4. Optional Remarks

The proof of the above theorem in [Pol90] is very readable (espcially if you ignore the section on pseudo-dimension which isn't critical). However, it's somewhat different from the "standard" proofs in other books that generally restrict attention to the case where δ is bounded by a constant. I prefer the extra generality above (since it comes in useful sometimes).

That said, the above theorem isn't really best possible. To get some intuition, let's think about the case where $\mathcal{F}$ is Euclidean. Then, in that case, the tails of Z are basically determined by δ, i.e. the maximal value of the process. This is in some ways analogous to Hoeffding's Inequality: the tails of a bounded random variable are bounded by the maximal size of the support. But we know that for random variables with small variance, Hoeffding's inequality can be quite loose. In the same way, for certain processes with small "variance", the the above theorem can be quite loose. You can make this heuristic argument rigorous by considering the special case when $\mathcal{F}$ is a singleton in Theorem 4.4 and comparing the given result to Bernstein's inequality.

[BLM13] present tighter bounds on the random variable Z that leverage "variance" information to help close the gap. The tightest known bounds are due to Talagrand (Talagrand's Inequality for the Suprema of the Empirical Process) but the proof is pretty dense, leverages a lot of special features, and (to my knowledge) requires $\hat{\boldsymbol{\xi}}_{1:n}$ to be i.i.d, not just independent.

References

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[GR21] Vishal Gupta and Paat Rusmevichientong. "Small-data, large-scale linear optimization with uncertain objectives". In: *Management Science* 67.1 (2021), pp. 220–241.

[Pol90] David Pollard. "Empirical Processes: Theory and Applications". In: Ims. 1990.