Borrowing Strength in Distributionally Robust Optimization via Hierarchical Dirichlet Processes

Nicola Bariletto[1], Khai Nguyen[1], and Nhat Ho[1]

[1]The University of Texas at Austin

May 23, 2024

Abstract

This paper presents a novel optimization framework to address key challenges presented by modern machine learning applications: High dimensionality, distributional uncertainty, and data heterogeneity. Our approach unifies regularized estimation, distributionally robust optimization (DRO), and hierarchical Bayesian modeling in a single data-driven criterion. By employing a hierarchical Dirichlet process (HDP) prior, the method effectively handles multi-source data, achieving regularization, distributional robustness, and borrowing strength across diverse yet related data-generating processes. We demonstrate the method's advantages by establishing theoretical performance guarantees and tractable Monte Carlo approximations based on Dirichlet process (DP) theory. Numerical experiments validate the framework's efficacy in improving and stabilizing both prediction and parameter estimation accuracy, showcasing its potential for application in complex data environments.

1 Introduction

A variety of problems in machine learning and statistics can be recast in terms of the following risk minimization problem:

$$\min_{\theta\in\Theta} \mathcal{R}_{p^\star}(\theta), \qquad \mathcal{R}_{p^\star}(\theta) := \mathbb{E}_{\xi\sim p^\star}[h(\theta,\xi)], \tag{1}$$

where $h : \Theta \times \Xi \to [0, K]$ is a loss function (e.g., the squared error [42, 13], negative log-likelihood [10], hinge [14], pinball [27], etc.) taking as input a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ and a data point $\xi \in \Xi \subseteq \mathbb{R}^p$.[1] Moreover, $p_\star$ is a data-generating distribution on the sample space Ξ (denoted as $p_\star \in \mathscr{P}_\Xi$), which is in general unknown. To make problem (1) feasible, one usually approximates $\mathcal{R}_{p_\star}(\theta)$ using a random sample $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$, a common choice being the empirical risk $\mathcal{R}_{p_\star}(\theta) \approx \mathcal{R}_{p_{\boldsymbol{\xi}}}(\theta) := N^{-1}\sum_{i=1}^N h(\theta, \xi_i)$.

While empirical risk minimization (ERM) remains popular, its effectiveness is frequently challenged by contemporary datasets. First, a common feature among such datasets is *high dimensionality*, which often requires imposing additional structure on the data-generating mechanism through regularization techniques. Secondly, the presence of complex generating processes coupled with limited sample sizes can lead to *distributional uncertainty*, prompting the need for strategies to mitigate it. Thirdly, data often display high levels of *heterogeneity*, for instance when observations are collected from diverse sources with related, but not identical data-generating processes.

[1]We assume Ξ is a measurable set and endow it with the relative Borel sigma algebra $\mathscr{B}(\Xi)$. Moreover, when relevant, we silently assume that maps with domain Ξ are measurable.

In the recent past, several strands of work have focused on addressing each of these challenges separately. First, popular strategies in the regularized estimation literature prescribe the penalization of the parameter complexity (e.g., as measured by different notions of a norm). In regression problems, this leads to well-known estimators like Ridge [24], LASSO [48], and the Elastic Net [55]. Most of these methods have also been shown to have a Bayesian interpretation [37, 29], leading to the proposal of further generalizations and improvements [49, 9]. Second, a large evolving body of literature on distributionally robust optimization (DRO) has provided several methods to deal with distributional uncertainty in a variety of optimization tasks (though not always with data-driven applications as the primary focus; see [39] for a recent review). The prevalent approach is min-max DRO (mM-DRO), whereby a worst-case risk criterion over an ambiguity set of plausible distributions is minimized [23, 2, 3, 15, 52, 16]. Recent notable contributions study mM-DRO problems in which the ambiguity set is defined as a Wasserstein ball of probability measures centered at the empirical distribution [35, 28]. Interestingly, while connections between these methods and regularization have been established [45, 12, 5, 44, 18], no obvious way to appropriately model heterogeneous data is available within the mM-DRO framework. Thirdly, a long-standing tradition in statistics, especially within the Bayesian community, has dealt with data heterogeneity, primarily via the development of complex hierarchical models [21, 20, 8]. These methods, however, mostly rely on the specification of a complete generative model, hindering their applicability to a variety of large-scale optimization-based methods in modern machine learning.

Contributions. In this work, we propose a novel optimization framework that allows to solve problem (1) in full generality and by *simultaneously* addressing the need for regularization, distributional robustness, and borrowing strength across heterogeneous data sources. In particular, assume we observe data coming from S different sources $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iN_i})$, $i = 1, \dots, S$, and denote by $\boldsymbol{\xi}^N$ the pooled sample of size $N = N_1 + \cdots + N_S$. For example, consider a collection of baseline health conditions measured on patients at S different hospitals, where we aim to assess the impact of such conditions on the efficacy of some specific treatment. To accommodate for the likely heterogeneity across groups, we propose replacing problem (1) with

$$\min_{\theta \in \Theta} V^s_{\boldsymbol{\xi}^N}(\theta), \qquad V^s_{\boldsymbol{\xi}^N}(\theta) := \int_{\mathscr{P}_\Xi} \phi\left(\mathcal{R}_p(\theta)\right) Q^s_N(\mathrm{d}p)$$

for each group $s = 1, \dots, S$. In the above expression, Q^s_N denotes the posterior distribution, conditional on the samples $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_S$, of the s-th group law p_s under a hierarchical Dirichlet process (HDP) prior on the vector of laws $(p_1, \dots, p_S)$, while $\phi : \mathbb{R} \to \mathbb{R}$ is a strictly increasing, convex, and twice continuously differentiable function [1]. As we show in the rest of the paper, this criterion ensures

1. *Regularization*, via the prior information encoded in the top-level base measure of the HDP model;
2. *Distributional robustness*, via the curvature of ϕ inducing ambiguity aversion [1, 26];
3. *Borrowing of strength* across data sources, via the hierarchical structure of the HDP model.

As we discuss hereafter, our proposal integrates insights from decision theory under ambiguity and recent advancements in Bayesian nonparametric statistics, with the goal of addressing generic data-driven optimization tasks involving regularization, managing distributional uncertainty, and accommodating data heterogeneity in a principled way. Importantly, throughout the article, we establish some key advantages of our criterion, among which (i) its favorable statistical properties in terms of finite-sample and asymptotic performance guarantees; (ii) the availability of tractable approximations that are amenable to

standard gradient-based optimization methods; and (iii) its ability to effectively borrow strength across heterogeneous groups, as well as to both improve and stabilize the out-of-sample performance of standard learning methods.

The rest of the paper is organized as follows. Section 2 discusses a general Bayesian optimization framework that naturally allows for the incorporation of borrowing strength and distributional robustness. In Section 3, we provide finite-sample and asymptotic performance guarantees for the proposed criterion. Section 4 proposes tractable approximations to our criterion based on Monte Carlo integration and well known representations of the Dirichlet process. Section 5 presents results from numerical experiments testing our HDP-based robust method. Finally, Section 6 concludes the paper. The supplementary material collects technical proofs (Appendix A), further background on the theoretical tools employed throughout the paper (Appendix B), and further details on the numerical experiments (Appendix C).

2 A Bayesian Approach to Optimization With Heterogeneous Data

Recall the multiple-source optimization problem introduced in Section 1: We have data $\boldsymbol{\xi}_s \in \Xi^{N_s}$ partitioned in distinct groups indexed $s = 1, \ldots, S$, and we are interested in solving

$$\min_{\theta \in \Theta} \mathcal{R}_{p_s^\star}(\theta)$$

for each group s, where $(p_1^\star, \ldots, p_S^\star)$ is the vector unknown data-generating processes for the observations in the S groups. For instance, going back to the scenario with data collected from patients at S hospitals, assume that each ξ_{sj} comprises measurements of baseline health conditions and an indicator of efficacy for a drug under examination. By choosing $h(\theta, \xi)$ as a regression (e.g., squared) loss function tailored to predict efficacy from baseline conditions, the optimization process described above outputs the best-fitting parameter vector $\theta_\star^s \in \arg\min_{\theta \in \Theta} \mathcal{R}_{p_s^\star}(\theta)$.

Two common strategies to tackle the problem in practice are as follows. First, one can assume that the S groups are completely homogeneous in distribution so that standard ERM would dictate solving the pooled optimization problem

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{s=1}^{S} \sum_{j=1}^{N_s} h(\theta, \xi_{sj}),$$

whereby each group is assigned the same optimal parameter value outputted by the above procedure. Second, one can assume that the laws of the distinct groups are completely unrelated to each other, so that, following the ERM paradigm, one would solve

$$\min_{\theta \in \Theta} \frac{1}{N_s} \sum_{j=1}^{N_s} h(\theta, \xi_{sj})$$

separately for each group $s = 1, \ldots, S$. But what if the laws governing the S samples are dependent, yet not identical? In the running example, although treatment variations across different facilities might introduce heterogeneity among patients sampled from different hospitals, some degree of similarity is nonetheless expected. Particularly in smaller sample sizes, it becomes crucial to leverage this partial homogeneity, as doing so opens the possibility for information borrowing and shrinkage across samples.

We now demonstrate that adopting a Bayesian approach facilitates the acknowledgment and efficient utilization of this partial homogeneity in a principled manner. Specifically, given the unknown nature of the group-specific laws, a well-established practice in Bayesian statistics and decision theory [41] is to

manage uncertainty by specifying a prior Q for the vector of laws $(p_1, \ldots, p_S)$.[2] In particular, we model the observations as partially exchangeable: For all $s = 1, \ldots, S$ and $j = 1, \ldots, N_s$,

$$\begin{aligned} \xi_{sj} \mid (p_1, \ldots, p_S) &\overset{\text{ind}}{\sim} p_s, \\ (p_1 \ldots, p_S) &\sim Q. \end{aligned}$$

This model implies that, while observations are exchangeable (i.e., distributionally homogeneous) within groups, they might feature different yet dependent laws across groups. Notice that the two extreme cases of identity in law and unconditional independence are accommodated by this construction by, respectively, concentrating the mass of Q on the "diagonal" $p_1 = \cdots = p_S$ or specifying it as a product distribution (i.e., forcing the p_s's to be unconditionally independent).

Given the above Bayesian model, a coherent way to approximate $\mathcal{R}_{p_s^\star}(\theta)$ is by posterior averaging, which leads to minimizing

$$\mathbb{E}_{p_s \sim Q_N^s}[\mathcal{R}_{p_s}(\theta)] = \mathbb{E}_{p_s \sim Q_N^s}[\mathbb{E}_{\xi \sim p_s}[h(\xi, \theta)]] \equiv \mathbb{E}_{\xi \sim p_s^N}[h(\theta, \xi)], \tag{2}$$

where Q_N^s is the marginal posterior distribution of p_s given $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_S$ and

$$p_s^N(\mathrm{d}\xi) := \int_{\mathscr{P}_\Xi} p_s(\mathrm{d}\xi) Q_s^N(\mathrm{d}p_s)$$

is the posterior predictive distribution for sample s.

2.1 The Dirichlet Process Prior

In the next Subsection, we specify the prior Q for the vector $(p_1, \ldots, p_S)$ as a hierarchical Dirichlet process [46, 47]. A key building block of the latter is the univariate Dirichlet process (DP), the cornerstone nonparametric prior for a single distribution p [17]. Intuitively, the DP is characterized by the following finite-dimensional distributions: $p \sim \mathrm{DP}(\alpha, P)$ implies that $(p(A_1), \ldots, p(A_k))$ follows a Dirichlet distribution with parameters $\alpha P(A_1), \ldots, \alpha P(A_k)$ for any finite measurable partition $\{A_1, \ldots, A_k\}$ of Ξ.[3] An important feature of the DP is its almost sure discreteness, which yields $p \overset{\mathrm{d}}{=} \sum_{j \geq 1} w_j \delta_{\xi_j}$ for appropriately defined sequences of random weights $(w_j)_{j \geq 1}$ and atoms $(\xi_j)_{j \geq 1}$. The DP is also conjugate with respect to exchangeable sampling, which implies

$$p \sim \mathrm{DP}(\alpha, P) \;\Rightarrow\; p \mid \xi_1, \ldots, \xi_n \sim \mathrm{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} P + \frac{n}{\alpha + n} \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}\right).$$

That is, conditional on the exchangeable sample $\xi_1, \ldots, \xi_n$, p is again a DP with updated concentration parameter $\alpha + n$ and centered at the predictive distribution $\frac{\alpha}{\alpha+n} P + \frac{n}{\alpha+n} \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$. The latter is a compromise between the prior guess P and the empirical distribution $\frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$, where relative weights are determined by α and n. The predictive distribution is also related to the celebrated Blackwell-MacQueen Pólya urn scheme (or Chinese restaurant process), which draws $(\xi_i)_{i \geq 1}$ with law $p \sim \mathrm{DP}(\alpha, p_0)$: Sample $\xi_1 \sim P$ and, for all $i > 1$ and $\ell < i$, set $\xi_i = \xi_\ell$ with probability $1/(\alpha + j - 1)$, else (with probability $\alpha/(\alpha + j - 1)$) sample $\xi_i \sim P$ [4].

[2] Note the distinction in notation: When considering the law of group s as a random probability within the Bayesian framework, we denote it as p_s, whereas when viewing it as the fixed but unknown data-generating process, we denote it as $p_s^\star$.

[3] Also, $\mathbb{E}[p(A)] = P(A)$ and $\mathbb{V}[p(A)] = (1 + \alpha)^{-1} P(A)[1 - P(A)]$ for any $A \in \mathscr{B}(\Xi)$, justifying the names of α and P. Moreover, under mild assumptions, the DP enjoys full topological support [34], as desirable in our setting.

2.2 The Hierarchical Dirichlet Process Prior

Recently, there has been increasing interest in the study of priors for dependent distributions, particularly within the realm of Bayesian nonparametric. Many innovative methods have emerged, aiming to devise dependent large-support priors that maintain tractability in the posterior and predictive structure [33, 36, 40, 30, 31, 6]. Given our objective of addressing data-driven optimization tasks under minimal distributional assumptions, we draw inspiration from this literature on nonparametric methods, specifically focusing on a seminal model in the field: The hierarchical Dirichlet process (HDP) prior [46, 47]. This model is specified hierarchically as follows:

$$\begin{aligned} p_s \mid p_0 &\overset{\text{iid}}{\sim} \mathrm{DP}(\alpha_s, p_0), \qquad s = 1, \dots, S, \\ p_0 &\sim \mathrm{DP}(\alpha_0, H). \end{aligned} \tag{3}$$

where H is a continuous distribution on $(\Xi, \mathscr{B}(\Xi))$. Intuitively, the HDP prior models the dependent distributions $p_1, \dots, p_S$ as conditionally independent DP's sharing the same centering distribution p_0, which is in turn assigned a DP prior. Dependence among each pair p_s and $p_{\tilde{s}}$, then, is cleverly induced by allowing them to share the same base measure p_0 a priori.

Further insight on the dependence induced by HDP priors is gained by considering their popular *Chinese restaurant franchise* (CRF) representation [47], which relies on the following metaphor: A franchise comprising S restaurants serves an infinite menu of dishes shared across restaurants and generated by the centering measure H. Each restaurant has an infinite number of tables, each serving a single dish and capable of seating an infinite number of customers. The top-level, franchise-wide DP controls the assignment of customers (observations) to dishes (across-groups clusters), while the bottom-level, restaurant-specific DP's controls the assignment of dishes to tables (within-sample latent clusters). This creates a two-stage clustering procedure that allows different group laws (restaurants) to borrow information from each other thanks to the common atoms (dishes).

Using the CRF construction (see Appendix B for a detailed derivation), one deduces that, by modeling the vector of unknown distributions $(p_1, \dots, p_S)$ as a HDP, the optimization problem (2) simplifies to the minimization of

$$\frac{N_s}{\alpha_s + N_s} \underbrace{\frac{1}{N_s} \sum_{j=1}^{N_s} h(\theta, \xi_{sj})}_{\text{(a)}} + \frac{\alpha_s}{\alpha_s + N_s} \left[\frac{N}{\alpha_0 + N} \underbrace{\frac{1}{N} \sum_{\ell=1}^{S} \sum_{j=1}^{N_\ell} h(\theta, \xi_{\ell j})}_{\text{(b)}} + \frac{\alpha_0}{\alpha_0 + N} \underbrace{\mathbb{E}_{\xi \sim H}[h(\theta, \xi)]}_{\text{(c)}} \right]. \tag{4}$$

The above formula has the following interpretation. For each sample s, we optimize a compromise between the within-group empirical risk (a) and some form of overall average risk (more on it later). This compromise is naturally guided by the sample size N_s and the sample-specific concentration parameter α_s. In particular, for a large sample size N_s, the within-sample empirical risk dominates. Moreover, the overall average risk component is itself a compromise between the overall across-group empirical risk (b) and the expected risk (c) with respect to the prior centering distribution H. Also in this case, the balance between these two components is controlled by the overall sample size N and the top-level concentration parameter α_0. Notice that the completely dependent (pooled optimization) case is obtained by taking the $\alpha_s \to \infty$ limit, while the unconditionally independent (separate optimization) case is obtained by taking the $\alpha_0 \to \infty$ limit. In both scenarios, the empirical risk is also regularized by a term depending on the centering distribution H. Figure 1 summarizes the key components of the HDP risk proposed in this work. In the next Subsection, we turn to discuss how to induce the last missing component: Distributional robustness.

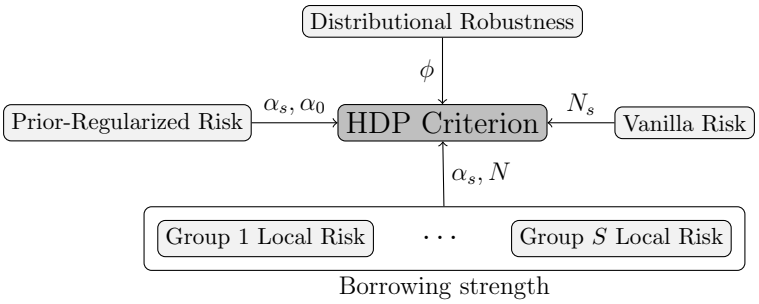


Figure 1: Key components of HDP-based optimization.

Example 1. *Assume each data point $\xi = (y, \boldsymbol{x}^\top)^\top \in \mathbb{R}^{p+1}$ consists of the measurement of p features $\boldsymbol{x}^\top = (x_1, \ldots, x_p)$ and a response y. If interested in linear prediction, one can choose $h(\theta, \xi) = (y - \boldsymbol{x}^\top\theta)^2$ for $\theta \in \mathbb{R}^p$, i.e., the classic quadratic loss. Moreover, assuming $H = \mathcal{N}(0, I)$, the ambiguity neutral criterion in the previous equation is equivalent to*

$$\sum_{j=1}^{N_s}(y_{sj} - \boldsymbol{x}_{sj}^\top\theta^s)^2 + \underbrace{\frac{\alpha_s}{\alpha_0 + N_0}\sum_{\ell=1}^{S}\sum_{j=1}^{N_\ell}(y_{\ell j} - \boldsymbol{x}_{\ell j}^\top\theta^s)^2}_{\text{Borrowing strength}} + \underbrace{\frac{\alpha_s\alpha_0}{\alpha_0 + N}\|\theta^s\|_2^2}_{\text{Ridge penalty}}.$$

*In other words, the HDP model with this specific choice of prior centering measure H yields **L_2-regularized least squares with borrowing of information** [1]: Estimation of θ^s is guided by information borrowed from the whole pooled sample $\boldsymbol{\xi}^N$ as well as the prior-induced L_2 penalization.*

2.3 Distributional Robustness via Smooth Ambiguity Aversion

Inspired by the seminal decision theory model of [26], [1] recently proposed a novel data-driven optimization framework, alternative to mM-DRO and based on the Bayesian approach adopted here, which effectively induces distributional robustness (see also [54], [53], [53], and [32] for previous related literature). Using our notation, the main idea is to transform the risk $\mathcal{R}_p(\theta)$ through a suitable deterministic function, before averaging it with respect to the posterior distribution of p. This basic method is easily adapted to our multiple-source setting by replacing the non-robust problem (2) with

$$\min_{\theta\in\Theta} V^s_{\boldsymbol{\xi}^N}(\theta), \qquad V^s_{\boldsymbol{\xi}^N}(\theta) := \int_{\mathscr{P}_\Xi} \phi\left(\mathcal{R}_p(\theta)\right) Q^s_N(\mathrm{d}p), \tag{5}$$

where $\phi : [0, K] \to \mathbb{R}$ is a convex, strictly increasing, and twice differentiable transformation. Criterion (5) incorporates (i) borrowing strength across groups, via the hierarchical structure of HDPs; (ii) regularization, via the prior information encoded in the base distribution H; and (iii) distributional robustness, via the curvature of ϕ. While (i) and (ii) follows from the above discussion on Equation (4) and Example 1, we illustrate (iii) as follows [1]. Consider the simple case when only two models, p^1 and p^2, are supported by a hypothetical posterior $Q = \frac{1}{2}\delta_{p^1} + \frac{1}{2}\delta_{p^2}$ (for the sake of the illustration, assume there is only one group, $S = 1$). Consider also two decisions θ_1 and θ_2 that, under p^1 and p^2, yield the expected risks marked on the horizontal axis of Figure 3 in Appendix B. While $\int \mathcal{R}_p(\theta_1)Q(\mathrm{d}p) = \int \mathcal{R}_p(\theta_2)Q(\mathrm{d}p) = \mathcal{R}^*$, the convexity of ϕ implies $\int \phi(\mathcal{R}_p(\theta_1))Q(\mathrm{d}p) < \int \phi(\mathcal{R}_p(\theta_2))Q(\mathrm{d}p)$. That is, although θ_1 and θ_2 yield the same loss in Q-expectation, the ambiguity-averse criterion favors θ_1 because it ensures less variability

across uncertain distributions p^1 and p^2.[4] In the data-driven context, this translates into a key robustness property that mitigates distributional uncertainty: Given limited amounts of data, one looks for a procedure yielding (i) good and (ii) not-too-variable out-of-sample performance across distributions over which there is uncertainty. Choosing a posterior Q with large enough support, then, ensures that this will hold for the underlying data-generating process as well.

3 Performance Guarantees

In this Section, we provide statistical guarantees on the performance of our robust optimization method. For each group $s = 1, \ldots, S$, denote by $\theta_N^s \in \arg\min_{\theta\in\Theta} V^s_{\boldsymbol{\xi}^N}$ a minimizer of the HDP criterion and recall $\theta_\star^s \in \arg\min_{\theta\in\Theta} \mathcal{R}_{p_s^\star}(\theta)$, our parameter of interest. In this setting, a natural measure of performance is the narrowness of the gap between $\mathcal{R}_{p_s^\star}(\theta_N^s)$ and $\mathcal{R}_{p_s^\star}(\theta_\star^s)$, and Lemma 2 is a first step towards establishing this type of guarantee.

Lemma 2. *Let ϕ be twice continuously differentiable on $(0, K)$, with $F_\phi := \sup_{t\in(0,K)} \phi'(t)$ and $S_\phi := \sup_{t\in(0,K)} \phi''(t) \geq 0$. Then*

$$\begin{aligned}
&\sup_{\theta\in\Theta} |V^s_{\boldsymbol{\xi}^N}(\theta) - \phi(\mathcal{R}_{p_s^\star}(\theta))| \\
&\qquad \leq \frac{N_s}{\alpha_s + N_s} F_\phi \sup_{\theta\in\Theta} |\mathcal{R}_{p_{\boldsymbol{\xi}_s}}(\theta) - \mathcal{R}_{p_s^\star}(\theta)| \\
&\qquad + \frac{\alpha_s}{\alpha_s + N_s} F_\phi \sup_{\theta\in\Theta} \left| \frac{N}{\alpha_0 + N} \mathcal{R}_{p_{\boldsymbol{\xi}^N}}(\theta) + \frac{\alpha_0}{\alpha_0 + N} \mathcal{R}_H(\theta) - \mathcal{R}_{p_s^\star}(\theta) \right| + \frac{K^2}{2} S_\phi.
\end{aligned}$$

Remark 3. *Lemma 2 provides insight into the benefits of regularization and borrowing strength as encoded in the criterion $V^s_{\boldsymbol{\xi}^N}(\theta)$. The result reveals that the maximum distance between the robust and the true criterion for sample s can be bounded by a weighted sum of (i) the maximum distance between the naive empirical risk criterion and the true one, (ii) the maximum distance between a mix of the regularized and empirical across-group criteria from the true criterion, and (iii) a term depending on the curvature of ϕ. In particular, this characterization implies that, if group s is similar in distribution to the other groups, the worst-case distance between our criterion and the ground truth can be improved through the borrowing of strength enabled by the proposed methodology. The same holds for the regularization term $\mathcal{R}_H(\theta)$, which can improve worst-case performance if it encodes accurate prior information on the data-generating process $p_s^\star$ (e.g., sparsity, correlation structure, and so on.).*

The next Proposition is useful because it reveals the possibility of obtaining finite-sample probabilistic performance certificates by establishing analogous guarantees for the naive empirical risk of each group s. The latter is a classic topic in modern statistical learning theory, which has produced a variety of techniques to ensure ERM convergence by imposing restrictions on the complexity of the function class $\mathscr{H} := \{h(\theta, \cdot) : \theta \in \Theta\}$, for instance by controlling its VC dimension, metric entropy, etc. We refer the reader to [51, 50] for an exhaustive treatment of the topic. We also highlight that such a straightforward transfer from classical theory to our methodology is a key dividend of the smoothness and tractability of the proposed criterion.

[4] Interestingly, [11] also showed a connection between this smooth ambiguity aversion model and mM-DRO. See Appendix B for further details.

Proposition 4. *For all $\delta > 0$,*

$$\begin{aligned}&\mathbb{P}[\phi(\mathcal{R}_{p_s^\star}(\theta_N^s)) - \phi(\mathcal{R}_{p_s^\star}(\theta_\star^s)) \leq \delta] \\ &\geq \mathbb{P}\left[\sup_{\theta\in\Theta} |\mathcal{R}_{p_{\boldsymbol{\xi}_s}}(\theta) - \mathcal{R}_{p_s^\star}(\theta)| \leq \frac{\alpha_s + N_s}{N_s}\left(\frac{\delta}{2F_\phi} - \frac{\alpha_s}{\alpha_s + N_i}K - \frac{K^2}{2}\frac{S_\phi}{F_\phi}\right)\right].\end{aligned}$$

Turning to asymptotic convergence results as the sample size N_s increases, note that finite-sample guarantees on $\sup_{\theta\in\Theta} |\mathcal{R}_{p_{\boldsymbol{\xi}_s}}(\theta) - \mathcal{R}_{p_s^\star}(\theta)|$ are usually of the form

$$\mathbb{P}\Big[\sup_{\theta\in\Theta} |\mathcal{R}_{p_{\boldsymbol{\xi}_s}}(\theta) - \mathcal{R}_{p_s^\star}(\theta)| \leq \delta\Big] \geq 1 - \eta_n,$$

with $\sum_{n=1}^\infty \eta_n < \infty$. This, in conjunction with the first Borel-Cantelli lemma, implies

$$\lim_{N_s\to\infty} \sup_{\theta\in\Theta} |\mathcal{R}_{p_{\boldsymbol{\xi}_s}}(\theta) - \mathcal{R}_{p_s^\star}(\theta)| = 0$$

almost surely, which we include as an assumption of the next Proposition proving convergence of optimal values to the true target. Moreover, we introduce a functional dependence of ϕ on n, and denote $\phi \equiv \phi_n$ accordingly.[5]

Proposition 5. *Retain the assumptions of Lemma 2 and, for all $s = 1, \ldots, S$, assume*

$$\lim_{N_s\to\infty} \sup_{\theta\in\Theta} \big|\mathcal{R}_{p_{\boldsymbol{\xi}_s}}(\theta) - \mathcal{R}_{p_s^\star}(\theta)\big| = 0$$

almost surely. Moreover, assume that ϕ_n satisfies (1) $\lim_{n\to\infty} S_{\phi_n} = 0$, (2) $\sup_{n\geq 1} M_{\phi_n} < \infty$, and (3) $\lim_{n\to\infty} \sup_{t\in[0,K]} |\phi_n(t) - t| = 0$. Then the next two limits hold almost surely for every sample $s = 1, \ldots, S$:

$$\lim_{N_s\to\infty} \mathcal{R}_{p_s^\star}(\theta_N^s) = \mathcal{R}_{p_s^\star}(\theta_\star^s), \qquad \lim_{N_s\to\infty} V_{\boldsymbol{\xi}^N}(\theta_N^s) = \mathcal{R}_{p_s^\star}(\theta_\star^s).$$

Finally, in the next Proposition, we prove convergence of the robust criterion optimizers to the target parameter depending on the true unknown data-generating mechanism.

Proposition 6. *Let $\theta \mapsto h(\theta, \xi)$ be continuous for all $\xi \in \Xi$ and $\lim_{N_s\to\infty} \mathcal{R}_{p_s^\star}(\theta_N^s) = \mathcal{R}_{p_s^\star}(\theta_\star^s)$ almost surely. Then, almost surely and for all samples $s = 1, \ldots, S$, $\lim_{N_s\to\infty} \theta_N^s = \bar{\theta}$ implies $\mathcal{R}_{p_s^\star}(\bar{\theta}) = \mathcal{R}_{p_s^\star}(\theta_\star^s)$.*

4 Monte Carlo Approximate Criterion

Due to the infinite dimensionality of the HDP marginal posterior $Q_N^i(\mathrm{d}p)$ and the non-linearity of the convex transformation ϕ, the integral defining $V_{\boldsymbol{\xi}^N}^i(\theta)$ in Equation (5) is analytically intractable. Hence, for practical implementation, we need to resort to suitable approximation schemes. The following result yields a key step in this direction.

[5] Note that the assumptions imposed on ϕ_n intuitively (and desirably) require that, as the sample size grows, the ambiguity aversion of the criterion vanishes (ϕ_{N_s} converges smoothly to the identity function). The assumptions are met by $\phi_n(t) = \beta_n \exp(t/\beta_n) - \beta_n$, with $\lim_{n\to\infty} \beta_n = \infty$, which from now on we silently assume to be our choice of ϕ.

Proposition 7 ([7], Theorems 9 and 10, Example 5)**.** *Assume the vector of laws* $(p_1, \dots, p_S)$ *is modeled as in Equation* (3) *and that the true laws* $(p_\star^1, \dots, p_\star^S)$ *are diffuse. Then*

$$p_0 \mid \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_S \sim \mathrm{DP}\left(\alpha_0 + N,\ \frac{N}{\alpha_0 + N}\frac{1}{N}\sum_{s=1}^{S}\sum_{j=1}^{N_s}\delta_{\xi_{sj}} + \frac{\alpha_0}{\alpha_0 + N}H\right),$$

$$p_s \mid p_0, \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_S \overset{\text{id}}{\sim} \mathrm{DP}\left(\alpha_s + N_s,\ \frac{N_s}{\alpha_s + N_s}\frac{1}{N_s}\sum_{j=1}^{N_s}\delta_{\xi_{sj}} + \frac{\alpha_s}{\alpha_s + N_s}p_0\right), \qquad s = 1, \dots, d.$$

Proposition 7 allows to set up a two-stage procedure to obtain an approximate Monte Carlo sample from the marginal posterior Q_N^s of p_s: First, simulate $p_0 \mid \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_S$, then simulate $p_s \mid p_0, \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_S$. Then, the sampled measure p_s (which, as a DP realization, is discrete) can be used to compute $\phi(\mathcal{R}_{p_s}(\theta))$, and a Monte Carlo average over many such samples approximate $V_N^s(\theta) = \int \phi(\mathcal{R}_{p_s}(\theta))Q_N^s(\mathrm{d}p_s)$. Moreover, this two-step procedure is simplified by the fact that, at both levels, we need to approximately simulate from DP distributions. This is possible by appealing to various DP constructions, e.g., via stick-breaking (SB) [43, 25] or multinomial-Dirichlet (MD) approximation [22]. In Appendix B, Algorithm 1 spells out the two-step simulation introduced above based on either SB or MD approximations of the DP, while Algorithms 2 and 3 provide details on SB and MD procedures.

5 Numerical Optimization and Experiments

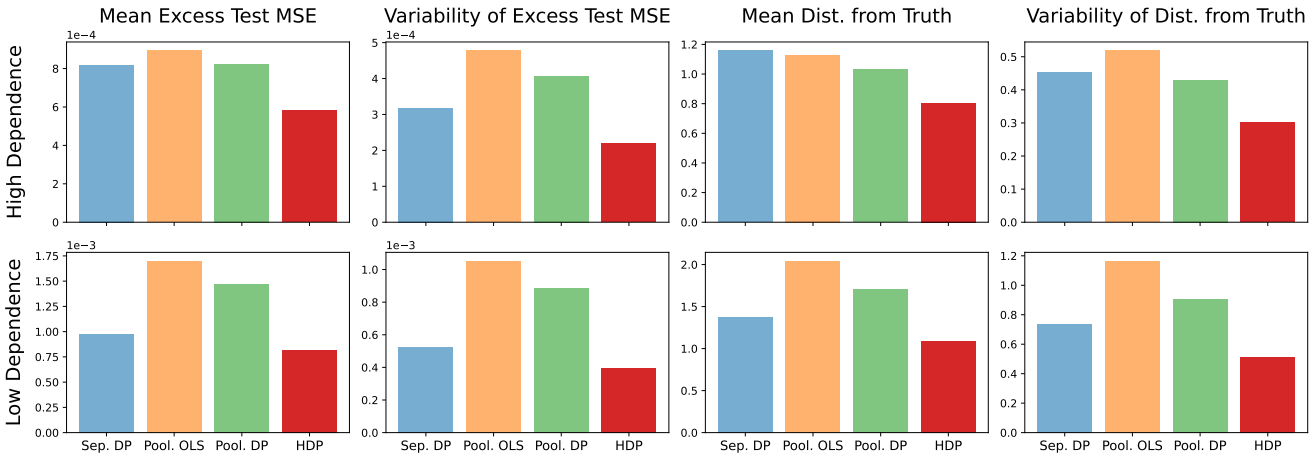
In this Section, we complement our theoretical analysis with empirical evidence on the benefits of our method. To the best of our knowledge, there is no obvious benchmark that allows for distributional robustness and borrowing strength at the same time. Therefore, in our numerical experiments with data coming from heterogeneous sources, we compare our method to (i) naive (non-robust) estimation procedures and (ii) the robust DP method of [1]. As neither alternative allows for borrowing strength, we implement both of them with two settings: By pooling the distinct sources into one single sample, and by keeping them entirely separate.

Numerical Optimization. Given the approximation strategies we proposed for the HDP criterion, in practice one ends up minimizing a function of the form

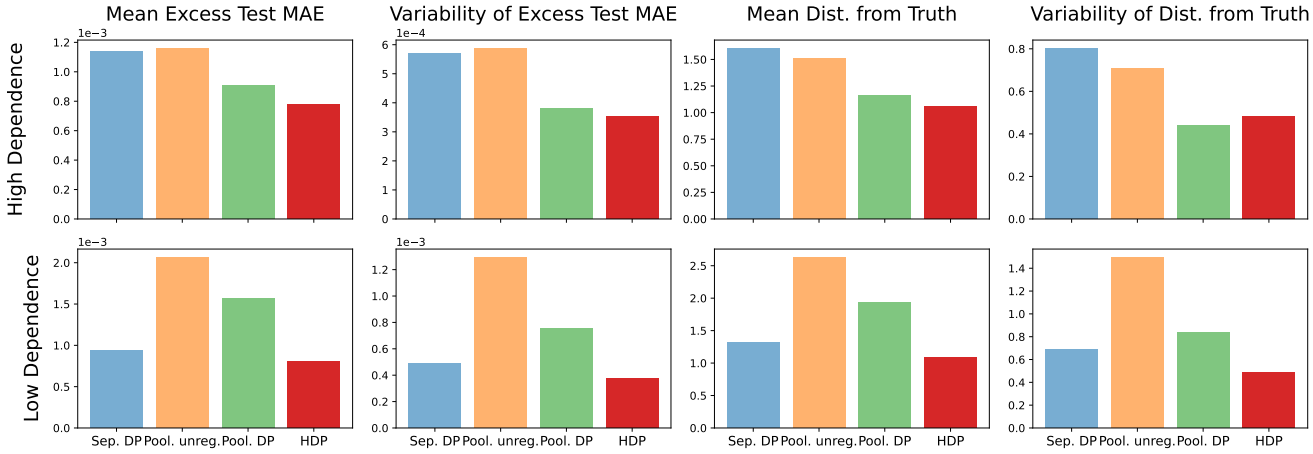
$$\hat{V}_s(\theta) = \frac{1}{M}\sum_{m=1}^{M}\phi\left(\sum_{j=1}^{T}p_j^m h(\theta, \xi_j^m)\right)$$

for each sample s. Under mild regularity assumptions on the loss function, $h(\theta, \xi)$, $\hat{V}_s$ is easily optimized via first-order methods. For instance, assuming we have access to the gradient $\nabla_\theta h(\theta, \xi)$ for all $\xi \in \Xi$, a simple yet scalable solution is to adopt a stochastic gradient descent (SGD) algorithm of the following type: At each iteration t, select a (possibly random) index $m_t \in \{1, \dots, M\}$, then perform a gradient-based update of the form

$$\theta^t = \theta^{t-1} - \eta_t \underbrace{\phi'\left(\sum_{j=1}^{T}p_j^{m_t}h(\theta^{t-1}, \xi_j^{m_t})\right)\sum_{j=1}^{T}p_j^{m_t}\nabla_\theta h(\theta^{t-1}, \xi_j^{m_t})}_{(\star)},$$



(a) Mean regression.



(b) Median regression.

Figure 2: Comparison of out-of-sample performance and estimation accuracy of different methods in high-dimensional regression experiments. The robust HDP method (in bright red) outperforms the others both in terms of the average and the variability of performance. Note: Distance from truth is measured as the square L_2 distance between the estimated coefficients and the data-generating ones.

where η_t is a pre-specified step size and $(\star)$ is an unbiased Monte Carlo estimate of the gradient of $\hat{V}_s$ evaluated at θ^{t-1}. This procedure highlights several appealing features of our method. First, the smoothness of the criterion opens the door for simple, off-the-shelf optimization procedures, which are not in general available for other DRO methods. Second, the convexity of h is easily seen to be inherited by $\hat{V}_s$, so that standard SGD convergence results for convex objectives hold [19]. Third, the form of the gradient allows to choose the truncation step T and the number of Monte Carlo samples M by interpreting them as the SGD minibatch size and the number of passes over the data, respectively. In fact, assume that the number of SGD iterations is chosen as a multiple of M and that m_t is chosen deterministically as follows: $m_0 = 1$ and $m_t = m_{t-1} + 1$. Then, the algorithm requires T gradient evaluations at each step, and it iterates N times over the whole (augmented) data. Finally, because ϕ is convex, ϕ' is increasing

so that ($\star$) can be interpreted as a form of *robustly weighted gradient*: The worse the current parameter value θ^{t-1} performs on the selected minibatch m_t, the more the procedure weights the corresponding gradient step.

Experiments. In the first experiment, we test the performance of the HDP robust criterion in a two-sample high-dimensional linear regression task, comparing it to OLS and robust DP estimation (both pooling samples and keeping them separate). In each simulation, we generate two size-100 samples of 95 features and a response, where the latter is linearly influenced by only 5 features. Moreover, the 5 non-zero coefficients are simulated at each iteration with positive correlation across samples, and we explore various degrees of dependence. Figure 2a shows the results of the study with 100 simulations, revealing that the robust HDP method outperforms, both in terms of out-of-sample risk and estimation accuracy, all of the other methods. Importantly, on top of doing better on average, HDP-robust estimation displays less variable performance. Figure 5 in Appendix C shows similar results when the degree of dependence among group laws is made even less or even more pronounced.

In the second simulation experiment, we test the performance of the HDP robust criterion in a two-sample high-dimensional median linear regression task, comparing it to the same baselines as above. Instead of recovering the conditional mean structure of the data-generating process, this method aims to reconstruct the conditional median of the response variable as a linear function of the features, which makes estimation more robust to outlier data points [27]. Using a data-generating process analogous to the first experiment, we test the ability of our HDP robust model to improve and stabilize performance when varying degrees of dependence are induced across heterogeneous groups. Figure 2b (and Figure 6 in Appendix C) show results in line with those of the linear (mean) regression above, with our robust HDP method in general outperforming the baselines on average and in terms of variability.

6 Discussion

In this paper, we put forward a data-driven optimization procedure that leverages hierarchical Dirichlet processes and a recently introduced Bayesian DRO framework to effectively induce regularization, distributional robustness, and borrowing strength among heterogeneous data sources. In particular, we provided performance guarantees of the proposed criterion, introduced Monte Carlo approximations that are easily optimized via gradient-based methods, and demonstrated the framework's efficacy through numerical experiments. While our results are promising, a few limitations are to be noted and leave room for further research on our method. For instance, additional testing on different loss functions, such as those used in deep learning architectures, would be beneficial to enhance applicability to a broader domain of learning algorithms. On the other hand, we believe in the possibility of obtaining more specific results (e.g., on rates of convergence and computationally efficient optimization solutions) for particular loss functions, such as the widely used squared loss for linear regression or, more generally, negative log-likelihoods of generalized linear models. Finally, the theoretical analysis may be made more general by relaxing some restrictive assumptions such as the boundedness of the loss function. While interesting, all of these points are beyond the scope of this work and thus left to future research.

References

[1] N. Bariletto and N. Ho. Bayesian Nonparametrics Meets Data-Driven Distributionally Robust Optimization. *arXiv preprint arXiv:2401.15771*, 2024.

[2] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[3] D. Bertsimas, X. V. Doan, K. Natarajan, and C.-P. Teo. Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research*, 35(3):580–602, 2010.

[4] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[5] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[6] F. Camerlenghi, D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodríguez. Latent nested nonparametric priors (with discussion). *Bayesian Analysis*, 14(4):1303–1356, 2019.

[7] F. Camerlenghi, A. Lijoi, P. Orbanz, and I. Prünster. Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92, 2019.

[8] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76, 2017.

[9] C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[10] G. Casella and R. Berger. *Statistical inference*. CRC Press, 2024.

[11] S. Cerreia-Vioglio, F. Maccheroni, M. Marinacci, and L. Montrucchio. Uncertainty averse preferences. *Journal of Economic Theory*, 146(4):1275–1330, 2011.

[12] R. Chen and I. C. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, 2018.

[13] R. Christensen. *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer, 2020.

[14] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[15] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[16] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[17] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

[18] R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 2022.

[19] G. Garrigos and R. M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

[20] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[21] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[22] S. Ghosal and A. Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[23] I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, 18(2):141–153, 1989.

[24] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[25] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[26] P. Klibanoff, M. Marinacci, and S. Mukerji. A smooth model of decision making under ambiguity. *Econometrica*, 73(6):1849–1892, 2005.

[27] R. Koenker. *Quantile regression*, volume 38. Cambridge University Press, 2005.

[28] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. INFORMS, 2019.

[29] Q. Li and N. Lin. The Bayesian elastic net. *Bayesian Analysis*, 5(1):151 – 170, 2010.

[30] A. Lijoi, B. Nipoti, and I. Prünster. Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20(3):1260–1291, 2014.

[31] A. Lijoi, B. Nipoti, and I. Prünster. Dependent mixture models: clustering and borrowing information. *Computational Statistics & Data Analysis*, 71:417–433, 2014.

[32] S. Lyddon, S. Walker, and C. C. Holmes. Nonparametric learning from Bayesian models with randomized objective functions. *Advances in Neural Information Processing Systems*, 31, 2018.

[33] S. N. MacEachern. Dependent Dirichlet processes. Department of Statistics, The Ohio State University, 2000.

[34] S. Majumdar. On topological support of Dirichlet prior. *Statistics & Probability Letters*, 15(5):385–388, 1992.

[35] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[36] P. Müller, F. Quintana, and G. Rosner. A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):735–749, 2004.

[37] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[39] H. Rahimian and S. Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022.

[40] A. Rodriguez, D. B. Dunson, and A. E. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

[41] L. J. Savage. *The foundations of statistics*. Courier Corporation, 1972.

[42] G. A. Seber and A. J. Lee. *Linear regression analysis*. John Wiley & Sons, 2003.

[43] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[44] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[45] S. Shafieezadeh Abadeh, P. M. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. *Advances in neural information processing systems*, 28, 2015.

[46] Y. Teh, M. Jordan, M. Beal, and D. Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. *Advances in neural information processing systems*, 17, 2004.

[47] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[48] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[49] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244, 2001.

[50] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

[51] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[52] W. Wiesemann, D. Kuhn, and B. Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[53] D. Wu, H. Zhu, and E. Zhou. A Bayesian risk approach to data-driven stochastic optimization: Formulations and asymptotics. *SIAM Journal on Optimization*, 28(2):1588–1612, 2018.

[54] E. Zhou and W. Xie. Simulation optimization when facing input uncertainty. In *2015 Winter Simulation Conference (WSC)*, pages 3714–3724. IEEE, 2015.

[55] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

Supplementary Material of "Borrowing Strength in Distributionally Robust Optimization via Hierarchical Dirichlet Processes"

This Supplement to "Borrowing Strength in Distributionally Robust Optimization via Hierarchical Dirichlet Processes" is organized as follows. In Appendix A, we collect the proofs of all results presented in the main text. In Appendix B, we provide further background on smooth ambiguity aversion, the hierarchical Dirichlet process prior, Dirichlet process representations, and algorithms. Finally, in Appendix C, we describe in detail the experiments presented in the paper.

A Technical Proofs

Proof of Lemma 2. First note that, by the stated assumptions, it follows from Taylor's theorem that

$$\phi(\mathcal{R}_p(\theta)) = \phi(\mathcal{R}_{p_i^\star}(\theta)) + \phi'(\mathcal{R}_{p_i^\star}(\theta))[\mathcal{R}_p(\theta) - \mathcal{R}_{p_i^\star}(\theta)] + \frac{\phi''(c_{p,\theta})}{2}[\mathcal{R}_p(\theta) - \mathcal{R}_{p_i^\star}(\theta)]^2$$

for all $p \in \mathscr{P}_\Xi$ and $\theta \in \Theta$ and for some $c_{p,\theta} \in [0, K]$. Then

$$\begin{aligned}
\sup_{\theta\in\Theta} |V^i_{\boldsymbol{\xi}^N}(\theta) - \phi(\mathcal{R}_{p_i^\star}(\theta))| &= \sup_{\theta\in\Theta} \left| \phi'(\mathcal{R}_{p_i^\star}(\theta)) \int_{\mathscr{P}_\Xi} [\mathcal{R}_p(\theta) - \mathcal{R}_{p_i^\star}(\theta)] Q^i_{\boldsymbol{\xi}^N}(\mathrm{d}p) \right. \\
&\quad \left. + \int_{\mathscr{P}_\Xi} \frac{\phi''(c_{p,\theta})}{2} [\mathcal{R}_p(\theta) - \mathcal{R}_{p_i^\star}(\theta)]^2 Q^i_{\boldsymbol{\xi}^N}(\mathrm{d}p) \right| \\
&\leq \frac{N_i}{\alpha_i + N_i} F_\phi \sup_{\theta\in\Theta} |\mathcal{R}_{p_{\boldsymbol{\xi}_i}}(\theta) - \mathcal{R}_{p_i^\star}(\theta)| \\
&\quad + \frac{\alpha_i}{\alpha_i + N_i} F_\phi \sup_{\theta\in\Theta} \left| \frac{N}{\alpha_0 + N} \mathcal{R}_{p_{\boldsymbol{\xi}^N}}(\theta) + \frac{\alpha_0}{\alpha_0 + N} \mathcal{R}_H(\theta) - \mathcal{R}_{p_i^\star}(\theta) \right| \\
&\quad + \frac{K^2}{2} S_\phi.
\end{aligned}$$

Proof of Proposition 4. Notice the following decomposition:

$$\begin{aligned}
&\underbrace{\phi(\mathcal{R}_{p_i^\star}(\theta^i_N)) - \phi(\mathcal{R}_{p_i^\star}(\theta^i_\star))}_{\geq 0} \qquad (6) \\
&= \phi(\mathcal{R}_{p_i^\star}(\theta^i_N)) - V^i_{\boldsymbol{\xi}^N}(\theta^i_N) + \underbrace{V^i_{\boldsymbol{\xi}^N}(\theta^i_N) - V^i_{\boldsymbol{\xi}^N}(\theta^i_\star)}_{\leq 0} + V^i_{\boldsymbol{\xi}^N}(\theta^i_\star) - \phi(\mathcal{R}_{p_i^\star}(\theta^i_\star)) \\
&\leq 2 \sup_{\theta\in\Theta} |V^i_{\boldsymbol{\xi}^N}(\theta) - \phi(\mathcal{R}_{p_i^\star}(\theta))|.
\end{aligned}$$

Then, Lemma 2 implies that, for all $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}[\phi(\mathcal{R}_{p_i^\star}(\theta^i_N)) - \phi(\mathcal{R}_{p_i^\star}(\theta^i_\star)) \leq \delta] \\
&\geq \mathbb{P}\left[\sup_{\theta\in\Theta} |V^i_{\boldsymbol{\xi}^N}(\theta) - \phi(\mathcal{R}_{p_i^\star}(\theta))| \leq \delta/2 \right] \\
&= \mathbb{P}\left[\sup_{\theta\in\Theta} |\mathcal{R}_{p_{\boldsymbol{\xi}_i}}(\theta) - \mathcal{R}_{p_i^\star}(\theta)| \leq \frac{\alpha_i + N_i}{N_i} \left(\frac{\delta}{2F_\phi} - \frac{\alpha_i}{\alpha_i + N_i} K - \frac{K^2}{2} \frac{S_\phi}{F_\phi} \right) \right].
\end{aligned}$$

Proof of Proposition 5. Since

$$\lim_{N_i\to\infty}\sup_{\theta\in\Theta}|\mathcal{R}_{p_{\boldsymbol{\xi}_i}}(\theta)-\mathcal{R}_{p_i^\star}(\theta)|=0$$

almost surely and given assumptions 1. and 2. on ϕ_n, by Lemma 2 we obtain

$$\lim_{N_i\to\infty}\sup_{\theta\in\Theta}|V_{\boldsymbol{\xi}^N}(\theta)-\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta))|=0$$

almost surely. Then, by decomposition (6),

$$\lim_{N_i\to\infty}\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_N^i))-\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_\star^i))=0$$

almost surely and

$$\lim_{N_i\to\infty}V_{\boldsymbol{\xi}^N}(\theta_N^i)-V_{\boldsymbol{\xi}^N}(\theta_\star^i)=0$$

almost surely. As a consequence,

$$\begin{aligned}
&\lim_{N_i\to\infty}|V_{\boldsymbol{\xi}^N}(\theta_N^i)-\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_\star^i))|\\
&\qquad\le\lim_{N_i\to\infty}\Big[|V_{\boldsymbol{\xi}^N}(\theta_N^i)-V_{\boldsymbol{\xi}^N}(\theta_\star^i)|+|V_{\boldsymbol{\xi}^N}(\theta_\star^i)-\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_\star^i))|\Big]\\
&\qquad\le\lim_{N_i\to\infty}\Big[|V_{\boldsymbol{\xi}^N}(\theta_N^i)-V_{\boldsymbol{\xi}^N}(\theta_\star^i)|+\sup_{\theta\in\Theta}|V_{\boldsymbol{\xi}^N}(\theta)-\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta))|\Big]\\
&\qquad=0
\end{aligned}$$

almost surely. Now recall assumption 3., i.e., the sequence $(\phi_n)_{n\ge1}$ converges uniformly to the identity map. Then, in light of the previous observations and by noticing that

$$\begin{aligned}
|\mathcal{R}_{p_i^\star}(\theta_N^i)-\mathcal{R}_{p_i^\star}(\theta_\star^i)|\le{}&|\mathcal{R}_{p_i^\star}(\theta_N^i)-\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_N^i))|+|\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_N^i))-\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_\star^i))|\\
&+|\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_\star^i))-\mathcal{R}_{p_i^\star}(\theta_\star^i)|
\end{aligned}$$

and

$$|V_{\boldsymbol{\xi}^N}(\theta_N^i)-\mathcal{R}_{p_i^\star}(\theta_\star^i)|\le|V_{\boldsymbol{\xi}^N}(\theta_N^i)-\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_\star^i))|+|\phi_{N_i}(\mathcal{R}_{p_i^\star}(\theta_\star^i))-\mathcal{R}_{p_i^\star}(\theta_\star^i)|,$$

the two desired almost sure limits follow:

$$\lim_{N_i\to\infty}\mathcal{R}_{p_i^\star}(\theta_N^i)=\mathcal{R}_{p_i^\star}(\theta_\star^i),\qquad\lim_{N_i\to\infty}V_{\boldsymbol{\xi}^N}(\theta_N^i)=\mathcal{R}_{p_i^\star}(\theta_\star^i).$$

Proof of Proposition 6. We have

$$\mathcal{R}_{p_i^\star}(\theta_\star^i)\le\mathcal{R}_{p_i^\star}(\bar\theta)=\mathbb{E}_{\xi\sim p_i^\star}\lim_{N_i\to\infty}h(\theta_N^i,\xi)=\lim_{N_i\to\infty}\mathcal{R}_{p_i^\star}(\theta_N^i)=\mathcal{R}_{p_i^\star}(\theta_\star^i)$$

almost surely, where the first equality follows from the continuity of $\theta\mapsto h(\theta,\xi)$ and the second one from the Dominated Convergence Theorem. Then, $\mathcal{R}_{p_i^\star}(\bar\theta)=\mathcal{R}_{p_i^\star}(\theta_\star^i)$ almost surely, proving the result.

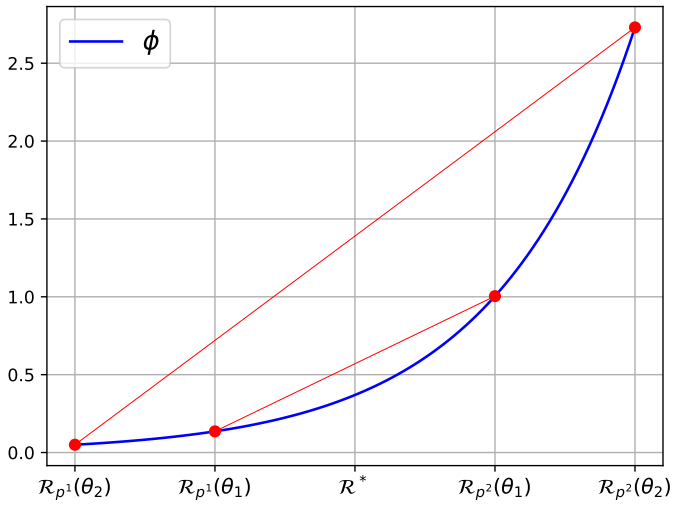


Figure 3: Graphical display of smooth ambiguity aversion at work [1]. Although θ_1 and θ_2 yield the same loss $\mathcal{R}^*$ in Q-expectation, the ambiguity averse criterion favors the less variable decision θ_1. Graphically, this is because the orange line connecting $\phi(\mathcal{R}_{p^1}(\theta_1))$ to $\phi(\mathcal{R}_{p^2}(\theta_1))$ lies (point-wise) below the line connecting $\phi(\mathcal{R}_{p^1}(\theta_2))$ to $\phi(\mathcal{R}_{p^2}(\theta_2))$.

B Further Background and Algorithms

B.1 Connections Between Smooth Ambiguity and mM-DRO

[11] showed that the smooth ambiguity aversion (SmAA) model belongs to a general class of ambiguity-averse preferences, which admit a common utility function representation. For SmAA preferences with $\phi(t) = \beta \exp(\beta^{-1}t) - \beta$ (with $\beta > 0$ and under additional technical assumptions), this representation implies the equivalence of problem

$$\min_{\theta\in\Theta} \int_{\mathscr{P}_\Xi} \phi(\mathcal{R}_p(\theta))Q(\mathrm{d}p)$$

with

$$\min_{\theta\in\Theta} \max_{P:P\ll Q_{\boldsymbol{\xi}^n}} \left\{\mathbb{E}_{p\sim P}[\mathcal{R}_p(\theta)] - \beta \mathrm{KL}(P\|Q)\right\},$$

where $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence and $\ll$ denotes absolute continuity. The above result further clarifies the mechanism through which distributional robustness is induced by ϕ: Intuitively, instead of directly averaging over $p \sim Q$, one computes a worst-case scenario w.r.t. the mixing measure, penalizing distributions that are further away from Q – the latter, which in our case coincides with a posterior distribution, acts as a reference probability measure. Moreover, in the limiting case $\beta \to 0$, the mM-DRO setup is recovered, with ambiguity set

$$\mathcal{C} = \left\{p \in \mathscr{P}_\Xi : \exists P \ll Q, p = \int_{\mathscr{P}_\Xi} qP(\mathrm{d}q)\right\}.$$

In the other limiting case $\beta \to \infty$ (with the convention $0 \cdot \infty = 0$), the ambiguity-neutral Bayesian criterion is instead recovered.

B.2 The Hierarchical Dirichlet Process Prior and its Chinese Restaurant Franchise Construction

The hierarchical Dirichlet process [46, 47] serves as a prior on a vector of dependent probability measures $(p_1, \dots, p_S)$, and is specified as follows:

$$\begin{aligned} \xi_{sj} \mid (p_1, \dots, p_S) &\overset{\text{ind}}{\sim} p_s, \qquad s = 1, \dots, S,\ j = 1, \dots, N_j \\ p_s \mid p_0 &\overset{\text{iid}}{\sim} \mathrm{DP}(\alpha_s, p_0), \qquad s = 1, \dots, S, \\ p_0 &\sim \mathrm{DP}(\alpha_0, H). \end{aligned}$$

where H is a continuous distribution on $(\Xi, \mathscr{B}(\Xi))$ and $\mathrm{DP}(\alpha, P)$ denotes the distribution of a Dirichlet process (DP) with concentration parameter $\alpha > 0$ and centering distribution P. This construction implies that the observations ξ_{sj} are *partially exchangeable*: Exchangeability (i.e., distributional invariance under index permutations) holds within each group s, but not necessarily across different groups $s \neq s'$, thus allowing for (partial) heterogeneity.

In order to derive expression (4), we need a characterization of the predictive distribution for group s, that is, $\mathbb{E}[p_s \mid \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_d]$. This is possible by leveraging the Chinese restaurant franchise construction of [46]. The metaphor goes as follows: A franchise of S restaurants shares dishes (unique values) drawn from a franchise-wide menu p_0, which is a weighted collection of dishes drawn from a DP with base measure H (the latter can be thought of as an infinitely rich source of recipes). Each restaurant has infinite capacity, meaning that it contains an infinite number of tables, each able to host an infinite number of customers – the only restriction is that customers seating at the same table will share the same dish. Now fix a restaurant s and assume we are given the configuration of the N_s customers $(\xi_{s1}, \dots, \xi_{sN_s})$ into T_s tables: Table θ_{s1} seats t_{s1} customers, table θ_{s2} seats t_{s2} customers, etc., with the obvious constraint $\sum_{j=1}^{T_s} t_{sj} = N_s$. Each table corresponds to an iid draw θ_{sj} from the franchise-level menu p_0, which we hold fixed for now. Then, because the HDP model places a $\mathrm{DP}(\alpha_s, p_0)$ at the level of restaurant s, the Chinese restaurant construction of the DP [4] implies that the next customer (observation) of restaurant s will be seated to table θ_{sj} with probability proportional to t_{sj}, or to a yet unoccupied table with probability proportional to α_s. In formulas,[6]

$$\xi_{sN_s+1} \mid \boldsymbol{t}_s, \boldsymbol{\theta}_s, p_0 \sim \sum_{j=1}^{T_s} \frac{t_{sj}}{N_s + \alpha_s} \delta_{\theta_{sj}} + \frac{\alpha_s}{N_s + \alpha_s} \delta_{\theta_{\text{new}}},$$

with $\theta_{\text{new}} \sim p_0 \mid \boldsymbol{t}_s, \boldsymbol{\theta}_s$. This procedure takes care of partitioning customers into tables within each restaurant. Notice that, because different tables can be assigned the same dish (Ξ value), the table configuration is only latent and instrumental to describe the predictive structure of the HDP.

Given the customer-table configurations of all restaurants, assume there are K distinct dishes being served in the whole franchise. That is, the tables θ_{sj} only feature K unique values $\xi_1^\star, \dots, \xi_K^\star$, with m_k tables serving dish $\xi_k^\star$ and, clearly, $\sum_{k=1}^K m_k = \sum_{s=1}^S T_s$. Then, because the HDP model places a $\mathrm{DP}(\alpha_0, H)$ prior on p_0, the Chinese restaurant process predictive construction of the DP implies[7]

$$\theta_{\text{new}} \sim \sum_{k=1}^K \frac{m_k}{\sum_{\ell=1}^K m_\ell + \alpha_0} \delta_{\xi_k^\star} + \frac{\alpha_0}{\sum_{\ell=1}^K m_\ell + \alpha_0} H.$$

[6] To keep the notation parsimonious, we identify the customer label ϕ with the table θ at which they sit.

[7] Again for parsimony of notation, we identify each table with the dish served at it.

See Figure 4 for a graphical illustration of this construction. Now recall our assumptions that the data is generated from a continuous distribution, implying that there are no ties among observations: $\xi_{sj} \neq \xi_{s'j'}$ for all $s, s' = 1, \ldots, S$, $j = 1, \ldots, N_s$, and $j' = 1, \ldots, N_{s'}$. This immediately implies that the only consistent table configuration in the Chinese restaurant franchise metaphor is the one in which all customers seat at a different table, each eating a different dish. In turn, this implies the predictive

$$\mathbb{E}[p_s \mid \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_d] = \frac{N_s}{\alpha_s + N_s} \frac{1}{N_s} \sum_{j=1}^{N_s} \delta_{\xi_{sj}} + \frac{\alpha_s}{\alpha_s + N_s} \left[\frac{N}{\alpha_0 + N} \frac{1}{N} \sum_{\ell=1}^{S} \sum_{j=1}^{N_\ell} \delta_{\xi_{\ell j}} + \frac{\alpha_0}{\alpha_0 + N} H \right],$$

yielding expression (4). This observation on the simplification of the table configuration in our continuous setting also yields the posterior characterization of Proposition 7. In fact, [7] provided a two-stage posterior characterization of the HDP that relies on the same type of latent table configuration appearing in the Chinese restaurant franchise. However, the no-ties assumptions in our setting makes it possible to simplify the characterization as in Proposition 7.

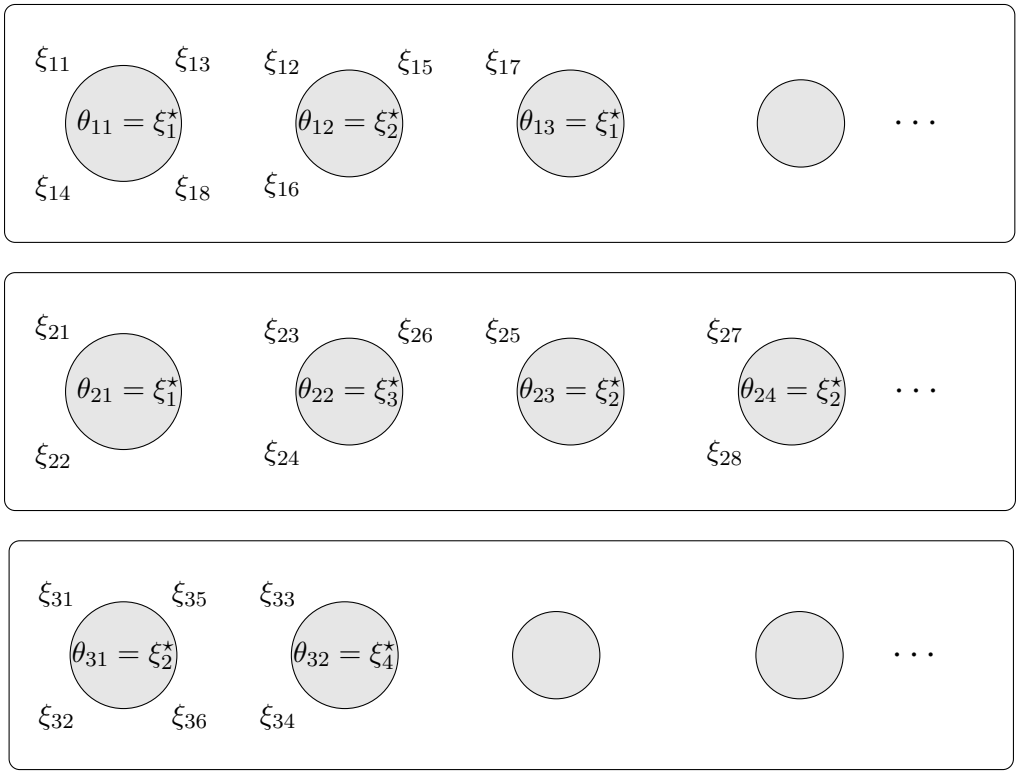


Figure 4: Illustration of the Chinese restaurant franchise construction of the HDP prior. In this example, there are $S = 3$ restaurants (represented by the rectangles) each hosting, at the current stage of the generative process, respectively 8, 8, and 6 customers. The restaurants seat their customers at 3, 4, and 2 tables, respectively, and a total number of $K = 4$ dishes $\xi_1^\star, \ldots, \xi_4^\star$ is served in the whole franchise.

Monte Carlo Approximation for the HDP Robust Criterion. Algorithm 1 summarizes the simulation strategy for the HDP robust criterion outlined in Section 4: Given (i) the posterior characterization of the HDP in the case with no ties among observations (see Proposition 7) and (ii) a method to simulate

from the DP (see the next two paragraphs for examples of such methods), one can repeatedly (a) simulate from the posterior of the top-level distribution p_0, and (b) given the realization of p_0, simulate from the posterior of the group-level distributions p_s. Finally, each group-specific criterion is approximated as a Monte Carlo average of the risks computed with respect to the simulated group-level distributions.

Algorithm 1 Monte Carlo Approximate HDP Criterion for Group s (HDP-MC$_s$)

Input: Data $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_d$, loss function h, function ϕ, concentration parameters α_s, α_0, top-level centering probability H, approximation type AP $\in$ {SB, MD}, truncation criteria T_s ($s \in \{1, \ldots, S\}$) and T_0, number of MC samples M

for $m = 1$ **to** M **do**

$(p_{0j}^m, \xi_{0j}^m)_{j=0}^{T_0} = \mathrm{AP}\big(\alpha_0 + N, \frac{\alpha_0}{\alpha_0+N} H + \frac{N}{\alpha_0+N} \frac{1}{N} \sum_{\ell=1}^{S} \sum_{j=1}^{N_\ell} \delta_{\xi_{\ell j}}, T_0\big)$

$\hat{p}_0^m = \sum_{j=0}^{T_0} p_{0j}^m \delta_{\xi_{0j}^m}$

$(p_{sj}^m, \xi_{sj}^m)_{j=0}^{T_i} = \mathrm{AP}\big(\alpha_s + N_s, \frac{\alpha_s}{\alpha_s+N_s} \hat{p}_0^m + \frac{N_s}{\alpha_s+N_s} \frac{1}{N_s} \sum_{j=1}^{N_s} \delta_{\xi_{sj}}, T_s\big)$

end for

Return: $\theta \mapsto M^{-1} \sum_{m=1}^{M} \phi\big(\sum_{j=0}^{T_s} p_{sj}^m h(\theta, \xi_{sj}^m)\big)$

Stick-Breaking Construction of the Dirichlet Process. [43] proved that Ferguson's 1973 Dirichlet process enjoys the following "stick-breaking" representation

$$p \sim \mathrm{DP}(\alpha, P) \implies p \overset{\mathrm{d}}{=} \sum_{j=1}^{\infty} p_j \delta_{x_j},$$

where

$$\begin{aligned} x_j &\overset{\mathrm{iid}}{\sim} P, \quad j = 1, 2, \ldots, \\ p_1 &= B_1, \\ p_j &= B_j \prod_{i=1}^{j-1} B_i, \quad j = 2, 3, \ldots, \\ B_j &\overset{\mathrm{iid}}{\sim} \mathrm{Beta}(1, \alpha), \quad j = 1, 2, \ldots \end{aligned}$$

The name of the procedure comes from the analogy with breaking a stick of length 1 into two pieces of length B_1 and $1 - B_1$, then the second piece into two sub-pieces of length $(1 - B_1)B_2$ and $(1 - B_1)(1 - B_2)$, and so on. Algorithm 2, then, presents a truncated version of the stick-breaking procedure, which stops at step T. The remaining portion of the stick is then allocated to one further atom drawn from the centering measure P.

Multinomial-Dirichlet Construction of the Dirichlet Process. Another finite-dimensional approximation of $p \sim \mathrm{DP}(\alpha, P)$ is $p_T = \sum_{j=1}^{T} p_j \delta_{x_j}$, with $x_j \overset{\mathrm{iid}}{\sim} P$ and $(p_1, \ldots, p_j) \sim \mathrm{Dirichlet}(T; \alpha/T, \ldots, \alpha/T)$. As $T \to \infty$, p_T approaches p [see Theorem 4.19 in 22]. Hence, one can approximately sample from a DP as in Algorithm 3. For all of our experiments (see Appendix C), we choose to simulate from the DPs at both levels of the hierarchy of the HDP via the multinomial-Dirichlet approximation. This is because, even for moderate T, this method tends to assign more balanced weights than the stick-breaking constructions, making practical optimization of the HDP robust criterion more stable.

Algorithm 2 Truncated Dirichlet Process Stick-Breaking Algorithm (SB)

Input: Concentration parameter α, centering probability P, truncation criterion $T \in \mathbb{N}$
Set $\prod_{k=1}^{0}(1 - B_k) \equiv 1$
for $j = 1$ **to** T **do**
  Draw $\xi_j \sim \pi$
  Draw $B_j \sim \text{Beta}(1, \alpha)$
  Set $p_j = B_j \prod_{k=1}^{j-1}(1 - B_k)$
end for
Draw $\xi_0 \sim \pi$
Set $p_0 = \prod_{k=1}^{T}(1 - B_k)$
Return: $(p_j, \xi_j)_{j=0}^{T}$

Algorithm 3 Truncated Dirichlet Process Multinomial-Dirichlet Algorithm (MD)

Input: Concentration parameter α, centering probability P, truncation criterion $T \in \mathbb{N}$
Initialize $\boldsymbol{p} = (p_1, \ldots, p_T) \in \mathbb{R}^T$
for $j = 1$ **to** T **do**
  Sample $\xi_j \sim P$
  Update $p_j \sim \text{Gamma}(\alpha/T, 1)$
end for
Normalize $\boldsymbol{p} = \frac{\boldsymbol{p}}{\sum_{j=1}^{n} p_j}$
Return: $(p_j, \xi_j)_{j=1}^{T}$

C Experiments

High-Dimensional Sparse Linear Regression Simulation Study. In this experiment, we take $h(\theta, \xi)$ to be the squared loss (as usual in linear regression tasks) and conduct simulations as follows. Denote by $s = 1, 2$ two distinct yet related samples consisting of $n = 100$ observations per sample, where each observation consists of $p = 95$ features, collected in a matrix $\boldsymbol{X}^s \in \mathbb{R}^{n \times p}$, and a target, collected in a vector $\boldsymbol{y}^s \in \mathbb{R}^n$. Observations are generated according to the following hierarchical model:

$$\begin{aligned}
\boldsymbol{y}^1 \mid \boldsymbol{X}^1, \boldsymbol{\beta}^1 &\sim N(\boldsymbol{X}^1\boldsymbol{\beta}^1, \sigma^2 I_n), \\
\boldsymbol{y}^2 \mid \boldsymbol{X}^2, \boldsymbol{\beta}^2 &\sim N(\boldsymbol{X}^2\boldsymbol{\beta}^2, \sigma^2 I_n), \\
\boldsymbol{X}_i^s &\overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma}_{p\times p}), \quad i = 1, \ldots, n,\ s = 1, 2,
\end{aligned}$$

where $\boldsymbol{\Sigma}_{p\times p}$ takes value 1 on its diagonal and 0.3 off-diagonal, and $\sigma = 0.5$. In order to induce dependence across samples 1 and 2, we generate the coefficient vectors $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$ as follows. First, to ensure sparsity, we set the last 90 coordinates of both vectors equal to 0. Second, we generate the first 5 coordinates as follows (denote by $\boldsymbol{\beta}_5^1$ and $\boldsymbol{\beta}_5^2$ the sub-vectors of active coefficients):

$$(\boldsymbol{\beta}_5^1, \boldsymbol{\beta}_5^2) \sim N(\boldsymbol{1}_{10}, c \cdot \boldsymbol{V}_{10\times 10}),$$

where $\boldsymbol{V}_{10\times 10}$ takes value 1 on its diagonal and 0.3 off-diagonal, while $c \in \{0.1, 0.2, 0.4, 0.5\}$ controls the degree of dependence among coefficients both across and within samples (the smaller c, the larger the correlation among coefficients). Finally, we take $H = N(\boldsymbol{0}, I_n)$ (prior centering distribution), $\phi(t) =$

Algorithm 4 Stochastic Gradient Descent Algorithm (SGD)

Input: Approximate criterion parameters $\{(p_j^m, \xi_j^m) : j = 1, \ldots, T,\ m = 1, \ldots, M\}$, loss function h, function ϕ step size schedule $(\eta_t)_{t \geq 1}$, starting value θ^0, number of iterations I
for $t = 1$ **to** I **do**
  Choose $m_t \in \{1, \ldots, M\}$
  Update $\theta^t = \theta^{t-1} - \eta_t \phi' \left(\sum_{j=1}^T p_j^{m_t} h(\theta^{t-1}, \xi_j^{m_t}) \right) \sum_{j=1}^T p_j^{m_t} \nabla_\theta h(\theta^{t-1}, \xi_j^{m_t})$
end for
Return: θ^I

$\exp(t) - 1$ (i.e., $\beta = 1$), $T_s = T_0 = 100$ (Multinomial-Dirichlet approximation steps at both levels of the hierarchy), $M = 300$ (number of MC samples from the HDP posterior), and set the SGD step size $\eta_t = 500/(100 + \sqrt{t})$. We run SGD until visually-inspected convergence, which takes around 2 seconds per run on our infrastructure (see Appendix D).

We first conduct 10 simulations through which we select the optimal concentration parameter values for the HDP procedure, the separate-samples DP procedure, and the pooled-samples DP procedure, across a grid of plausible values. The selection is performed by fitting the models on training samples generated as above, then computing the out-of-sample risk on 10,000 additional simulated test observations. Using the optimized parameter values, we run 100 more simulations and, for each of these, compute (1) the excess[8] out-of-sample loss (mean squared error) and (2) the squared L_2 distance between the estimated and true coefficient vectors. Figure 5, which shows results for $c \in \{0.1, 0.5\}$, confirms the results of Figure 2a in the main body, which shows results for $c \in \{0.2, 0.4\}$: Even in the presence of more extreme dependence structures (i.e., very low dependence or very high dependence, based on the more extreme values of c), the HDP method does better both on average and in terms of reduced variability, compared to the alternatives.[9]

High-Dimensional Sparse Linear Median Regression Simulation Study. In this experiment, we take h to be the pinball loss with quantile parameter 0.5, $h(\theta, \xi) = |y - \theta^\top \boldsymbol{x}|$, which aims to recover the conditional median of the response variable y given the feature vector $\boldsymbol{x}$. In terms of generative process and simulation setting, we keep everything as in the previous experiment on linear (mean) regression. We also keep the DP and HDP robust criterion parameters as before, and set the SGD step size $\eta_t = 500/(100 + \sqrt{t})$. We run SGD until visually-inspected convergence, which takes around 12 seconds per run on our infrastructure (see Appendix D).

Figures 2b and 6 report the results from 100 simulation after 10 initial ones for parameter selection (analogously to the previous experiment). The Figure 2b in the main text reports results for low and high across-groups dependence regimes ($c \in \{0.2, 0.4\}$), and in Figure 6 these regimes are taken to even larger extremes. In both cases, the qualitative conclusions highlighted for the linear regression experiment hold as well: The HDP-robust method, compared to the baseline robust DP and naive ERM estimation strategies, is effective at (i) borrowing information across groups to an optimal extent, and (ii) managing distributional uncertainty by reducing performance variability.

[8]The word "excess" refers to the risk computed at the true underlying parameter, which, in the context of simulation studies like this, we obviously have access to.

[9]Notice that the Figures do not report results for the separate-samples OLS procedures. This is because, both in terms of average performance and its variability, this method performs worse than the others by one order of magnitude, and including it in the plots would distort relative comparisons among the other methods. Nevertheless, we refer the reader to our code for results on this procedure as well.

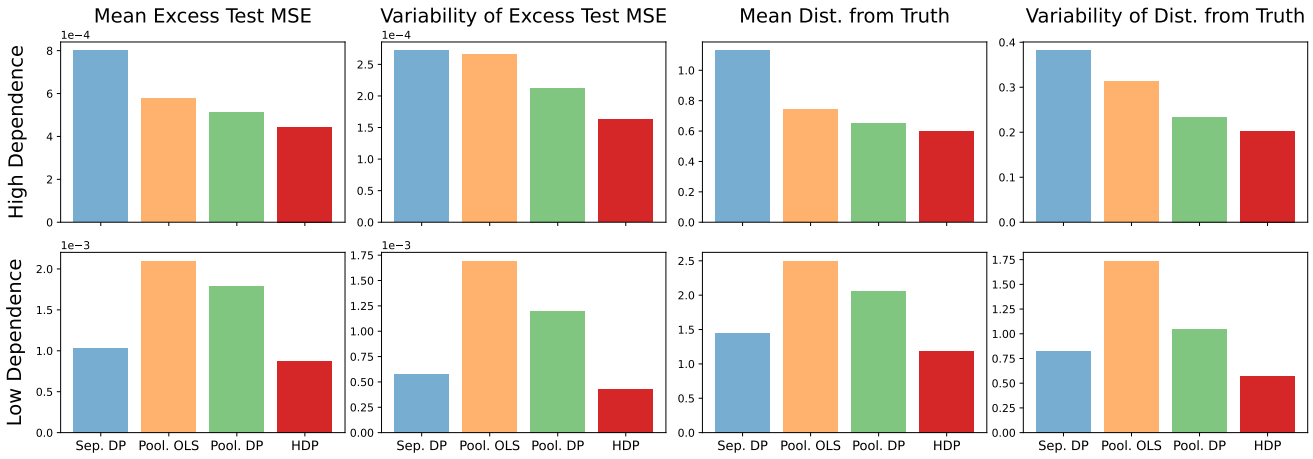


Figure 5: Out-of-sample performance and estimation accuracy of different methods in the high-dimensional linear regression experiment. The HDP method (in bright red) outperforms the others both in terms of the average and the variability of performance. Distance from truth is measured as the square L_2 distance between the estimated coefficients and the data-generating ones. Note: OLS estimation was performed using the Python library `scikit-learn` [38].

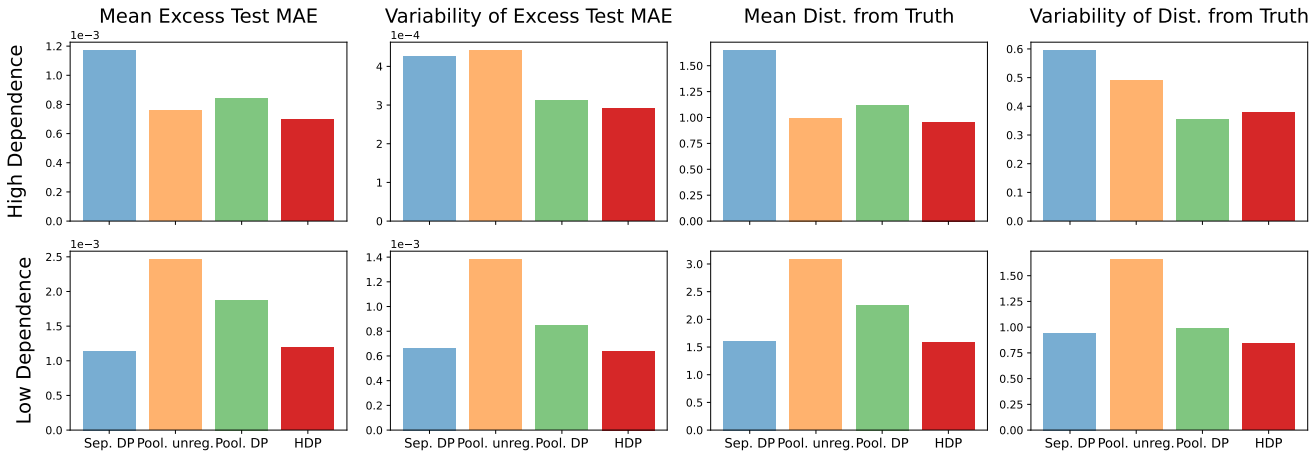


Figure 6: Out-of-sample performance and estimation accuracy of different methods in the high-dimensional linear median regression experiment for dependence parameter values $c \in \{0.1, 0.5\}$. The HDP method (in bright red) outperforms the others in terms of performance variability, and does as well on average. Distance from truth is measured as the square L_2 distance between the estimated coefficients and the data-generating ones. Note: unregularized estimation was performed using the Python library `scikit-learn` [38].

D Computational Infrastructure

All experiments were performed on a desktop with 12th Gen Intel(R) Core(TM) i9-12900H, 2500 Mhz, 14 Core(s), 20 Logical Processor(s) and 32.0 GB RAM.