Robust Stochastic Optimization: Learning the Tails

John C. Duchi
Hongseok Namkoong
Stanford University, 350 Serra Mall, Stanford CA 94305 USA

Abstract

We develop and analyze a robust stochastic optimization framework that learns a solution which is robust to perturbations in the underlying distribution. We formulate a convex procedure for the finite sample approximation and provide statistical guarantees, showing that the finite sample problem concentrates around the robustified population objective. The robust solutions optimize performance on the tails of the input distribution instead of the average performance. Simulation experiments show that robust solutions outperform the empirical risk minimizer under adversarial perturbations in the underlying distribution by optimizing the performance on the tails of the input distribution.

1. Introduction

In many modern applications of machine learning, the data generating distribution under which the samples came from changes at test time. This may be due to inherent non-stationarity in the underlying system, unexpected hardware failures or interactions with other potentially adversarial agents. In this paper, we consider a learning framework that is explicitly robust to changes in the underlying distribution, optimizing performance on the longer-tailed inputs instead of the average performance. In situations requiring high reliability—guaranteed levels of performance nearly all the time instead of on average—such robustness is a natural goal.

To that end, we develop and analyze the following robust stochastic optimization problem:

\[
\minimize_{x \in X} \sup_{P \in P_0} \left\{ \mathbb{E}_P[\ell(x; \xi)] : D_f(P \| P_0) \leq \rho \right\}. \tag{1}
\]

In the problem (1), \(P_0\) is a fixed base distribution on the set \(\Xi\); for each \(\xi \in \Xi\), the function \(x \mapsto \ell(x; \xi)\) is closed and convex; and \(D_f(P \| Q)\) denotes the \(f\)-divergence between \(P\) and \(Q\),

\[
D_f(P \| Q) := \int f \left( \frac{dP}{dQ} \right) dQ
\]

and \(f : \mathbb{R}_+ \rightarrow \mathbb{R}\) is a closed convex function satisfying \(f(1) = 0\) and \(f(t) = +\infty\) for any \(t < 0\).

Taking the worst-case approach to ensure robustness is almost classical at this point (see, for example, the book of Ben-Tal et al. (2009)). Recently, a number of authors extended such methods to settings where one observes a sample \(\xi_1, \ldots, \xi_n\) and constructs an uncertainty set over the data directly (Delage & Ye, 2010; Wang et al., 2013; Ben-Tal et al., 2013; Bertsimas et al., 2013; 2014). Robustification over the observed sample can be viewed as a regularization—for example, Xu et al. (2009a;b) note this equivalence for support vector machines and Lasso regression for certain uncertainty sets. The viewpoint taken in this paper is somewhat different in that we consider the population problem (1) affected by an adversarial (and potentially infinite-dimensional) perturbation of the underlying distribution \(P_0\). We propose computationally tractable procedures and provide statistical guarantees for solving this robust problem (1).

2. Convex Reformulation

We begin by providing a duality argument to transform problem (1) into a finite-dimensional problem. Consider the inner supremum in problem (1). Define the likelihood ratio \(L(\xi) := dP(\xi)/dP_0(\xi)\), in which case the supremum problem is equivalent to maximizing

\[
\sup_{\xi \in \Xi} \left\{ \int \ell(x; \xi) L(\xi) dP_0(\xi) : \int f(L(\xi)) dP_0(\xi) \leq \rho, \mathbb{E}_{P_0}[L(\xi)] = 1 \right\}. \tag{2}
\]

The following result then follows, where \(f^*(s) = \sup_{t \geq 0} \{ts - f(t)\}\) denotes the convex conjugate of \(f\):
Lemma 1. Assume that \( \rho > 0 \). Then

\[
\sup_{P \in P_0} \left\{ \mathbb{E}_P[\ell(x; \xi)] : D_f(P||P_0) \leq \rho \right\} = \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ \int_{\Xi} \lambda f^* \left( \frac{\ell(x; \xi) - \eta}{\lambda} \right) dP_0(\xi) + \lambda \rho + \eta \right\}.
\]

Moreover, if at \( x \in \mathcal{X} \) the supremum on the left hand side is finite, there are finite \( \lambda(x) \geq 0 \) and \( \eta(x) \in \mathbb{R} \) attaining the infimum on the right hand side.

In view of Lemma 1, we can solve the empirical version of the dual formulation to obtain an finite sample approximation to the robust problem (1). Let \( \hat{P}_n \) denote the empirical distribution given the sample. Then if \( \ell(\cdot; \xi) \) is convex for \( P_0 \)-almost all \( \xi \in \Xi \), the optimization problem

\[
\sup_{P \in P_n} \left\{ \mathbb{E}_{\hat{P}_n}[\ell(x; \xi)] : D_f(P||\hat{P}_n) \leq \rho \right\} \quad \text{(3)}
\]

is jointly convex in \( (x, \lambda, \eta) \) (Lemma 1 applies to any distribution \( P \), including the empirical distribution \( \hat{P}_n \)).

For concreteness, consider the Cressie-Read family (4) of divergences (Cressie & Read, 1984). Parameterized by \( k \), this family is given by

\[
f_k(t) = \frac{t^k - kt + k - 1}{k(k - 1)} \quad \text{(4)}
\]

where it is implicit that \( f_k(t) = \infty \) for \( t < 0 \). The above expression is defined as the continuous limit when \( k \to 0, 1 \) (resp. KL-divergence and log likelihood). After a calculation minimizing \( \lambda \) out in Lemma 1, the dual reformulation for the Cressie-Read family is

\[
\inf_{x \in \mathcal{X}, \eta} \left( 1 + k(k - 1)\rho \right) \left( \mathbb{E}_{P_0}[\ell(x; \xi) - \eta]_{+}^{k} \right)^{1/k} + \eta
\]

where \( 1/k + 1/k_{+} = 1 \). Roughly speaking, we are penalizing large upward deviations of \( \ell(x; \xi) \) above a certain quantile determined by \( \rho \). We may impose a stronger notion of robustness by picking \( k \) closer to 1. The form (5) can be also be interpreted as minimizing a coherent tail risk measure (Shapiro et al., 2009) instead of a simple average.

3. Convergence of the Empirical Minimizer

We now consider the convergence of the finite sample problem (3) to the population quantity. For fixed \( x \in \mathcal{X} \), we study the concentration of the expression (3). Throughout this section, we assume that for any \( x \in \mathcal{X} \) and \( \xi \in \Xi \), we have \( \ell(x; \xi) \in [0, 1] \). Suppressing the dependence on \( x \in \mathcal{X} \), we denote \( Z = \ell(x; \xi) \). We further restrict attention to the Cressie-Read divergence family (4) with \( k > 1 \). Define the function \( g_n : \mathbb{R}^n \to \mathbb{R} \) via

\[
g_n(Z) = \sup_{p : D_f(P||P_n) \leq \rho} \langle p, Z \rangle.
\]

We would first like to show that \( g_n(Z) \) concentrates around its mean. Noting that \( \mathbb{R}^n \ni Z \mapsto g_n(Z) \) is convex as a function of \( Z \), we consider the Lipschitzian properties of \( g \) with respect to the \( \ell_2 \)-norm. The subgradients of \( g_n \) are

\[
p_* \in \arg \max_{p : D_f(P||P_n) \leq \rho} \langle p, Z \rangle.
\]

that is, all vectors \( p_* \in \mathcal{P}_{n,k} \) attaining the supremum (Hiriart-Urruty & Lemaréchal, 1993, Corollary 4.4.4). From the \( f \)-divergence constraint, we have that

\[
\sum_{i=1}^{n} p_i^k \leq n^{1-k}(k(k-1)\rho + 1).
\]

When \( k \in (1, 2] \), since \( ||-||_k \geq ||-||_2 \), we have \( ||p||_2 \leq n^{-\frac{1}{k^2}}(k(k-1)\rho + 1)\frac{\hat{t}}{2} \). For \( k \geq 2 \), we have from Holder’s inequality that

\[
\sum_{i=1}^{n} p_i^2 \leq \left( \sum_{i=1}^{n} p_i^k \right)^{\frac{2}{k}} n^{-\frac{2}{k^2}} \leq n^{-1}(k(k-1)\rho + 1)\frac{\hat{t}}{2}.
\]

In particular, we obtain that \( g \) is Lipschitz with respect to the \( \ell_2 \)-norm with Lipschitz constant \( n^{-\frac{1}{k^2}}(k(k-1)\rho + 1)\frac{\hat{t}}{2} \) when \( k \in (1, 2] \) and \( n^{-\frac{1}{k^2}}(k(k-1)\rho + 1)\frac{\hat{t}}{2} \) when \( k \geq 2 \). We now recall a standard concentration inequality for Lipschitz convex functions of bounded random variables.

Lemma 2 (Boucheron et al. 2013, Theorem 6.10). Let \( h : \mathbb{R}^n \to \mathbb{R} \) be convex or concave and \( L \)-Lipschitz with respect to the \( \ell_2 \)-norm. Let \( Z_i \) be independent random variables with \( Z_i \in [a, b] \) and \( Z = Z_1^n \). For \( t \geq 0 \),

\[
\mathbb{P}(|h(Z) - \mathbb{E}[h(Z)]| \geq t) \leq 2 \exp \left( -\frac{t^2}{2L^2(b-a)^2} \right).
\]

Using Lemma 2, we obtain via our Lipschitz argument that for \( Z_i \in [0, 1] \),

\[
\mathbb{P} \left( |g_n(Z) - \mathbb{E}_{P_0}[g_n(Z)]| \geq t \right) \leq 2 \exp \left( -\min \left\{ n, \frac{n^{\frac{2}{k^2}}}{2} \right\} t^2 \right) \quad \text{(7)}
\]

To show concentration for the empirical version of our robust problem (1), it remains to show that

\[
\mathbb{E}_{P_0}[g_n(Z)] - \sup_{p : D_f(P||P_n) \leq \rho} \mathbb{E}_{P}[Z] \to 0
\]
at an appropriate rate. First, we note that by the representation (3), we have

\[ \mathbb{E}_{P_0} [g_n(Z)] = \mathbb{E}_{P_0} \left[ \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \lambda f^* \left( \frac{Z_i - \eta}{\lambda} \right) + \lambda \rho + \eta \right\} \right] \]

\[ \leq \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \mathbb{E}_{P_0} \left[ \frac{1}{n} \sum_{i=1}^{n} \lambda f^* \left( \frac{Z_i - \eta}{\lambda} \right) + \lambda \rho + \eta \right] \]

\[ = \sup_{\rho \in \mathcal{P}} \mathbb{E}_{P_0} [Z_1], \quad (8) \]

so we always have the upper that the expectation of the supremum is less than the supremum of expectations. To see the other direction, the following proposition provides a good lower bound. We omit the proof for brevity.

**Lemma 3.** Let \( k_* \in [1, \infty) \) and let \( Y_i \) be an i.i.d. sequence of random variables satisfying \( \mathbb{E}|Y_i^{2k_*}| \leq C k_* \mathbb{E}|Y_i|^{k_*} \) for some \( C \in \mathbb{R}_+ \). For any \( k_* \in [1, \infty) \), we have

\[ \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} |Y_i|^{k_*} \right)^{\frac{1}{k_*}} \right] \geq \mathbb{E}|Y_i^{k_*}|^{\frac{1}{k_*}} - 2C^{\frac{k}{k_*}} \frac{k_* - 1}{k_*} \cdot \left\{ \begin{aligned} & n^{\frac{k}{k_*}} \quad \text{if } k_* \geq 2 \\ & n^{-\frac{k}{k_*}} \quad \text{if } k_* < 2. \end{aligned} \right. \]

Now, we apply Lemma 3 to the representation (5) to obtain the final result.

**Theorem 1.** Let \( Z_i \in [0, 1] \) and \( k \in (1, \infty) \). Then, for \( g_n(Z) \) defined as in (6), we have the following concentration of the robust objective: for \( \delta > 0 \), with probability at least \( 1 - \delta \),

\[ \left| g_n(Z) - \sup_{P: P_J(P) \leq \rho} \mathbb{E}_{P}[Z_1] \right| \leq \max \left\{ n^{\frac{k}{k_*} - 1}, n^{-\frac{k}{k_*}} \right\} \left( \frac{2}{k} \sqrt{2(k(k-1)\rho + 1) \log \frac{2}{\delta}} \right). \]

**Proof** Define

\[ h(\eta; Z) := (1 + k(k-1)\rho) \left( [Z - \eta]_{+}^{k_*} \right)^{\frac{1}{k_*}} + \eta \]

for notational convenience. For \( \eta < 0 \), \( \mathbb{E}[h(\cdot; Z)] \) is differentiable, and its derivative is

\[ -(1 + k(k-1)\rho) \left( \mathbb{E}[Z - \eta]_{+}^{k_*} \right)^{-\frac{1}{k_*}} \left( \mathbb{E}[Z - \eta]_{+}^{k_*} - 1 \right) \leq -(1 + k(k-1)\rho) + 1 = -k(k-1)\rho \]

where the inequality followed from Jensen’s inequality. Since \( \mathbb{E}[h(\cdot; Z)] \) is decreasing in \( \eta \) on \( (-\infty, 0) \), we have

\[ g_n(Z) = \inf_{\eta \geq 0} \mathbb{E}_{P_0} [h(\cdot; Z)] \]

and similarly, the population counterpart \( g = \inf_{\eta \geq 0} \mathbb{E}_{P_0} [h(\cdot; Z)] \). Denote by \( \eta_0^* \) and \( \eta^* \) the minimizers of the respective optimization problems.

Noting that \( [Z - \eta]_+ \in [0, 1] \) for \( \eta \geq 0 \), we have

\[ \mathbb{E}_{P_0} \left[ \left( \mathbb{E}_{P_0} [Z - \eta]_{+}^{k_*} \right)^{\frac{1}{k_*}} \right] \geq \left( \mathbb{E}_{P_0} [Z - \eta]_{+}^{k_*} \right)^{\frac{1}{k_*}} - \frac{2}{k} \max \left\{ n^{\frac{k}{k_*}}, n^{-\frac{k}{k_*}} \right\} \]

for all \( \eta \geq 0 \) by Lemma 3. Thus

\[ 0 \leq g - \mathbb{E}_{P_0} [g_n(Z)] \leq g - \mathbb{E}_{P_0} [h(\eta_0^*; Z)] + \mathbb{E}_{P_0} [h(\eta_0^*; Z) - g_n(Z)] \leq \mathbb{E}_{P_0} \left[ h(\eta_0^*; Z) - \mathbb{E}_{P_0} [h(\eta_0^*; Z)] \right] \leq \frac{2}{k} \max \left\{ n^{\frac{k}{k_*}}, n^{-\frac{k}{k_*}} \right\} \]

where the first inequality follows (8), the second from the definition of \( \eta_0^* \), and the result then follows from (7). \( \square \)

From Theorem 1, we obtain a uniform concentration result in \( \mathcal{X} \) via covering number arguments (Boucheron et al., 2013). The rate of concentration deteriorates when the desired level of robustness is high (\( k \) close to 1 and \( \rho \) large); we believe these rates are unimprovable.

**4. Simulations**

In this section, we present three simulation experiments as a proof of concept: the hinge loss, \( \ell_1 \)-regression and \( \ell_2 \)-regression. We simulate on synthetic data where inputs \( a \in \mathbb{R}^d \) are randomly generated and the label/explanatory variable \( b \) is set according to a true classifier/regressor plus noise. To test whether our robust problem (1) performs well under adversarial perturbations in the data generating distribution \( P_0 \), we generate the test data by rotating the true classifier/regressor to the opposite direction incrementally. We observe that the robust solution has comparable average performance to that of the empirical risk minimizer but outperforms the ERM on the tails of the input distribution.

The loss differences are plotted in Figures 1, 2, 3. We define the risk \( R(x) = \mathbb{E}_{P_0} [\ell(x; \xi)] \). We solve for different values of \( k \) and \( \rho \) and observe that the solutions are indeed more robust when \( k \) is closer to 1 and \( \rho \) is large.

**5. Conclusion**

We have given a worst-case formulation for stochastic optimization as well as convex procedures for computing its finite sample approximations, relating them to long-tailed performance. We showed concentration of finite sample variants around their population counterparts, giving rates dependent on the desired level of robustness. Simulation
 experiments confirm the robust solutions performance under perturbations of the underlying data distribution.

We are actively investigating efficient procedures for solving the empirical problem (3). Scaling this problem as the sample size $n$ and dimension $d$ grow is imperative for application of the robust problem (1) in large-scale systems. We also plan to perform extensive real-world experimentation to see the practical effects of robustification.

References


