Direct Data-Driven Methods for Decision Making under Uncertainty

Junjie Qin JQIN@STANFORD.EDU
Institute for Computational and Mathematical Engineering, Stanford, CA 94305 USA

1. Introduction

We are constantly making decisions under uncertainty. A widely used formulation for decision making under uncertainty can be summarized by the following optimization program

$$u^\star = \operatorname*{argmin}_{u\in\mathcal{U}} \mathbb{E}_{\mathbb{P}}[f(u,X)], \tag{1}$$

where the future cost f depends both on the decision $u \in \mathcal{U}$ as well as the outcome of uncertain events, represented by a random variable $X \in \mathcal{X}$. Here the random variable X follows a distribution $\mathbb{P}$ which is assumed to be known in order to form the expectation in problem (1). Examples includes making an inventory decision with uncertainty future demand, purchasing stocks with uncertain information about the future price, and deciding which PhD programs to choose to go with uncertainty in research directions, advisors and funding opportunities.

In practice, the distribution $\mathbb{P}$ is never readily available. Instead, practitioners often resort to the following two step procedure:

$$\underset{(S_n)}{\text{Data}} \xrightarrow[\text{methods}]{\text{learning}} \underset{(\mathbb{P})}{\text{Model of uncertainty}} \xrightarrow[\text{optimization}]{\text{stochastic}} \underset{(u^\star)}{\text{Decision}}$$

that is, given a historical data set S_n, one would first apply certain machine learning algorithms to obtain representations of the data, usually in the form of point prediction or parameters for the distribution $\mathbb{P}$, and then use these representations to form the stochastic optimization program (1) which in turn would lead to an "optimal" decision. This procedure has many known issues. First of all, off-the-shelf learning algorithms are derived using loss functions that are mathematically convenient, but may have nothing to do with the actual economic costs regarding the decision that one is making. For instance, commonly used linear regression algorithm assumes a quadratic error loss, whereas the actual cost function may be a different function such as a piecewise linear function in the inventory control example. As such, the learning step in the above procedure is sub-optimal with respect to the true economic cost. Furthermore, as the stochastic optimization step sees the model of uncertainty which is a summary of the data instead of the data itself, it may make assumptions inconsistent with assumptions made in the learning step. An example would be using Gaussian error assumption in the stochastic optimization step whereas a ℓ_1-type regularization was used in the learning step (which implies the assumption that the error follows the Laplace distribution). Last but not the least, as different assumptions and approximations may be used in each step of the two-step procedure, it is usually very hard to theoretically gauge the performance of this two-step procedure.

An attractive proposal is to integrate these two steps. A couple of existing papers have explored this idea within particular application domains. Liyanage & Shanthikumar (2005) proposed the concept of operational statistics that drives the optimal estimator for the newsvendor ordering target under uncertainty. Their methods assume the functional form of the distribution for the uncertainty (e.g. the net demand follows an exponential distribution) and equivalence-type argument to reduce the hypothesis class of all possible estimators. Kao et al. (2009) considers the setting that the results of a regression are used for solving decision problems whose cost function is an arbitrary convex quadratic function. It is shown in the paper that by using a convex combination of the results produced by ordinary least squares and empirical minimization, the actual cost generated by their algorithm is significantly lower than directly using the results of ordinary least squares. A few important question still remains open. (i) What is a suitable general formulation for deriving methods that directly identify the optimal decision from the data? (ii) Are there algorithms that have provably guaranteed performances under this general formulation?

This paper makes progresses in addressing these questions. In particular, we adopt and modify the framework of *statistical decision theory*, which is the root of all modern learning algorithms, to incorporate the actual economic costs in Section 2. This leads to a risk minimization problem which again depends on the unknown underlying data generating distribution. We show that empirical risk minimization can lead to a class of efficient algorithms whose performance can be theoretically analyzed. Section 3 then applies these general consideration to a specific problem on dispatching

energy generators to meet uncertain demand in the context of renewable integration. We then propose algorithm that follows empirical risk minimization paradigm as well as variations of it. Theoretical guarantees are then derived followed by empirical test results with data from BPA.

2. Formulation

Given a family of distributions $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$, and samples $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathbb{P}_\theta$, the goal of statistical decision theory is to identify a good procedure δ that maps the data $X = (X_1, \ldots, X_n)$ to a decision $d \in \mathcal{D}$ that has small risk. The notion of risk is defined as the expected value of a loss function $L(\theta, d)$, i.e.,

$$R(\theta, \delta) = \mathbb{E}_{\mathbb{P}_\theta}[L(\theta, \delta(X))],$$

where the loss function $L(\theta, d)$ assigns preference to decisions given the value of the model parameter θ and $d = \delta(X)$. This theory is usually used for the purpose of estimating the parameter θ itself of some known function of θ and so the resulting procedure δ is often called an estimator.

To cast our problem into the framework of decision theory, we use the following specification or modification

- We make no assumption on the form of the distribution $\mathbb{P}_\theta$. That is, instead of say assuming the uncertainty follows a normal distribution and then estimating its mean and variance, we allow arbitrary type of distribution in the family $\mathcal{P}$.[1] In a sense, this forms a nonparametric estimation problem in the classical terminology. As such, we drop the index θ and use $\mathbb{P}$ itself to refer to an arbitrary member of the family of the unknown distribution $\mathcal{P}$.

- We set the loss function to be the true economic cost of the problem, that is

  $$L(X_{n+1}, d) = f(d, X_{n+1}),$$

  where $X_{n+1} \sim \mathbb{P}$ denotes the unobserved uncertainty in when we are making the decision (i.e. X in (1)). Notice that this modifies the conventional loss function which is a deterministic function of the unknown parameter θ and decision d to a function that depends on the random realization of X_{n+1}. The resulting risk function is

  $$\mathbb{E}_{\mathbb{P}} f(u, X_{n+1}),$$

  where we modified the notation from estimator δ to decision u.

[1] We would still have to make implicit technical assumptions such as the mean of the loss function exists.

The remaining problem is to design procedures that maps the data to a good decision u that minimizes the risk. The challenge is that since $\mathbb{P}$ is unknown, it is in general impossible to find procedures that are *uniformly optimal* with respect to all possible members $\mathbb{P} \in \mathcal{P}$. Thus the bulk of the classical decision theory concerns about alternative notations of optimality, which leads to uniformly minimal risk *unbiased* estimators, uniformly minimal risk *equivariant* estimators, optimal *Bayes* estimators, and *minimax* estimators (cf. Lehmann & Casella (1998) for a good treatment of this subject). However, none of the above optimality criteria permits universal procedures to derive algorithms which could identify the optimal decision. That is, calculation has to be done based on the particularly assumed distribution, and for different distributions the methods and results vary significantly.

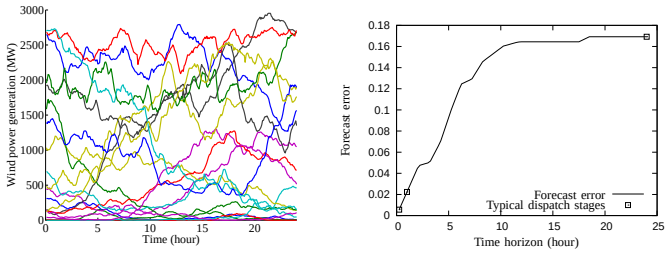
We propose, instead, to minimize the empirical risk $(1/n)\sum_{i=1}^n f(u, X_i)$ within a pre-determined hypothesis[2] class that contains functional forms for u. This would certainly generate efficient algorithms if for examples the cost function is convex and the hypothesis class is linear. In more complex situations, non-convex optimization procedures may be deployed to identify the optimal hypothesis in the hypothesis class. We will show by an application in the next section that it is possible to prove theoretical performance guarantees for the resulting procedures which suggests that the sub-optimality with respect to the true risk R (which is defined using the unknown distribution $\mathbb{P}$) is bounded with large probability and the bound approaches to zero with the number of samples increases to infinity. The application consists of a specific choice of the cost function for a practical situation, and the hypothesis class. But both the proposed algorithms and the performance analysis could be generalized to other cost functions and hypothesis classes.

3. Application: Electric Power Dispatch for Renewable Integration

The rest of this paper concerns with an application that is of an increasing importance both in the United States and around the globe. As global warming becomes a growing consensus, many countries around the world are pushing deeper renewable penetration into their energy generation portfolio and their electric power grids. This results in significant challenges in the operation of the grids as renewables are intrinsically variable, i.e., they are intermittent, uncontrollable and random. Figure 1 (a) depicts the wind power generation in a BPA region over 20 days, and Figure 1 (b) gives common percentage forecast errors for the

[2] This use of the term of hypothesis follows leaning theory instead of the literature of hypothesis testing, although they are closely related.

wind with different forecast horizons. It is evident that as the forecast errors for the wind is significant at day-ahead (around 16% in Figure 1(b)), an explicit modeling of its uncertainty and its impact is necessary. See (Qin et al., 2013a) and (Qin et al., 2013b) for more backgrounds.



(a) Wind generation for 20 days (b) Forecast error (%) for wind

Figure 1. Renewable is variable and difficult to forecast.

3.1. Problem Statement

The problem that we are concerning is the dispatch of conventional generators, which are slow and have to be notified and scheduled at 24 hours ahead of the delivery period, in a grid with substantial amount of the renewables in the system. This is a challenging problem in the sense that when the decision regarding the conventional generator is made, no precise information about the net demand, defined as the actual demand minus the renewable generation is available. We can formulate this problem as a specific instance of the general decision-making under uncertainty problem that we discussed in the previous sections. In particular, if the distribution for the net demand is $\mathbb{P}$, then one would like to identify the optimal amount to be scheduled for the conventional generators by solving the following stochastic optimization

$$u^\star = \operatorname*{argmin}_{u \in \mathcal{U}} \mathbb{E}_{\mathbb{P}}[cu + q(D-u)_+], \tag{2}$$

where u is the dispatched slow generator at day-ahead with cost per unit generation being c dollars, $D \in [D^{\min}, D^{\max}]$ is the future net demand for the system, $(D-u)_+ = \max(D-u, 0)$ is the unserved demand with per unit penalty being q dollars. In California, the average per unit cost for slow generators (i.e. c) is around \$50 per MW, whereas the penalty for each unit of unserved demand (i.e. q, also referred to as value of lost load) is in the range of $\$500 \sim 2000$ per MW. So we assume that $q > c$.

As a matter of practice, the actual distribution for the net demand D is unknown. Instead, we have observation of the net demand over historical hours, together with other relevant data that we refer to as features. For instance, for a historical hour i, we may have records of the net demand D_i and a vector of features X_i with its entries recording the temperature, wind speed, and other relevant information about hour i, as well as some nonlinear transformations of the some of the features available. Denote the historical data set by $S_n = \{(X_1, D_1), \ldots, (X_n, D_n)\}$, where $X_i \in \mathbb{R}^m$. We are also given features regarding the delivery hour which we are making a dispatch decision about, denoted as X_{n+1}. Now our goal is to find a mapping from X_{n+1} to u that has small risk using data S_n. As in Section 2, the true population risk is defined using economic cost of the system, which coincides with the objective function of the stochastic optimization (2), i.e.,

$$R = \mathbb{E}_{\mathbb{P}}[cu + q(D-u)_+].$$

3.2. Algorithms

As the true population risk cannot be evaluated without the knowledge about the underlying distribution $\mathbb{P}$, we minimize the empirical risk instead. Furthermore, we restrict ourselves to the hypothesis class of linear functions of the feature X_{n+1} for tractability. Note that this does not reduce the generality of our procedure a lot because as mentioned before we can always add nonlinear transformed features into the original list of features to capture any nonlinear effects. Within the linear hypothesis class, the dispatch decision u can be represented by a weight vector $w \in \mathbb{R}^m$, i.e., $u = \sum_{j=1}^m w^j X_{n+1}^j$.

We proposes three algorithms listed as follows:

- Algorithm 1: Empirical Risk Minimization (ERM)

  $$\min_w \frac{1}{n} \sum_{i=1}^n cw^\top X_i + q(D_i - w^\top X_i)_+$$

  with its solution denoted by w^{ERM}.

- Algorithm 2: ERM with Least Squares Regularization

  $$\min_w \frac{1}{n} \sum_{i=1}^n cw^\top X_i + q(D_i - w^\top X_i)_+ + \lambda_{\mathrm{LS}} \|D - Xw\|_2,$$

  with its solution denoted by $w^{\mathrm{ERM+LS}}$.

- Algorithm 3: ERM with ℓ_2 Regularization

  $$\min_w \frac{1}{n} \sum_{i=1}^n cw^\top X_i + q(D_i - w^\top X_i)_+ + \lambda_2 \|w\|_2,$$

  with its solution denoted by $w^{\mathrm{ERM}+\ell_2}$.

Here the first algorithm is directly minimizes the empirical risk $\hat{R} = \frac{1}{n}\sum_{i=1}^n cw^\top X_i + q(D_i - w^\top X_i)_+$ within the hypothesis class. Algorithm 2 is proposed for the case that some of the underlying data-generating features may not

be included in the model. If the missing components can be approximated with normal distribution (by the virtue of the central limit theorem), the least square regularization would improve ERM. The last algorithm is proposed for the case that certain automatic feature selection is needed: If the number of features is large, the ℓ_2 regularization in Algorithm 3 would be useful to reduce over-fitting.

3.3. Performance Guarantees

We have the following guarantees on the performance of w^{ERM}.

Theorem 1 (Uniform Convergence Bound)**.** *For i.i.d. data $(X_1, D_1), \ldots, (X_n, D_n)$, suppose that we restrict to the weights satisfying $\|w\|_2 \leq W^{\max}$ and suppose that $\mathbb{E}\|X_i\|^2 \leq (X^{\max})^2$, then with probability at least $1-\delta$, the excess risk of ERM is bounded as*

$$|R(w^{\mathrm{ERM}}) - R(w^\star)| \leq \frac{4(q-c)W^{\max}X^{\max}}{\sqrt{n}} + \sqrt{\frac{2\log(2/\delta)}{n}},$$

where $w^\star$ is the minimizer of the population risk R.

This result suggest that with probability $1-\delta$, the excess risk of ERM diminishes as $O(1/\sqrt{n})$, i.e., as the number of samples grows to infinity, the result of ERM is near-optimal in the hypothesis class with large probability. One unsatisfactory fact regarding the previous result is that it does not show how the algorithm performs as the number of feature m changes. Using tools from algorithmic stability theory, the next results bounds the generalization error of Algorithm 1 and explicit shows the dependence on the number of features.

Theorem 2 (Algorithmic Stability Bound)**.** *Under assumptions of Theorem 1 and w.l.o.g. assume $|D^{\max}| > |D^{\min}|$, with probability at least $1-\delta$, the generalization error of ERM is bounded as*

$$\begin{aligned}&|R(w^{\mathrm{ERM}}) - \hat{R}(w^{\mathrm{ERM}})| \\ \leq& 2\gamma(q-c)D^{\max}\frac{m}{n} + (q-c)D^{\max}(4\gamma m+1)\sqrt{\frac{\log(2/\delta)}{2n}}\end{aligned}$$

where $\gamma = (q-c)/c$.

This result shows that generalization error scales as $O(m/\sqrt{n})$ (note the second term in the bound) so that when we have a large number of features, the sample size has to grow much faster to ensure the same risk bound. Note that the generalization error is different from the excess risk and is only the difference of between the population risk and empirical risk both evaluated at the point produced by the ERM algorithm. As the algorithmic stability theory concerns the output of the particular algorithm, it does not bound the risk difference with the true population risk minimizer. However, we would expect the error bounds in Theorem 2 to be informative as well, because in general whenever the empirical risk and population risk are close enough (which is ensured by the bound given), the distance between their minimizers should not be far.

3.4. Numerical Results

We test all three algorithms with one year of hourly wind and demand data from BPA (`http://transmission.bpa.gov/business/operations/wind/`). The three proposed algorithms are compared with a benchmark algorithm which is separated estimation and optimization (SEO), that is the two-step procedure discussed in Section 1. Here since for the deterministic case of (2) the optimal solution is clearly $u = D$, the SEO reduces to performs a least square regression for the demand. We have tested three sets of features. The first set of features consists of the last net demand observation and hour of the day, where the hour of the day is used to capture seasonality. The second set of features consists of the net demand over the last 24 hours and hour of the day. The last set of features includes the net demand over the last 24 hours, order statistics[3] of the net demand and hour of the day. All three cases contains a constant feature representing the intercept, so that the number of features for these three cases are $m = 3$, $m = 26$, and $m = 50$, respectively. The tests are conducted in a rolling horizon fashion: for each hour in which a dispatch decision has to be made, the data corresponding to the past n hours are used as the sample data set. The same procedure is repeatedly tested for all $N - n$ hours, where $N = 24 \times 365 = 8760$ is the total number of hours in the data set.[4]

The results for our experiments are shown in Table 1 in the form of the percentage cost reduction compared to the benchmark algorithm. From the results in the table, we can observe that with small number of features ($m = 3$), the cost saving of ERM increases slowly as the number of samples increases. While with more features ($m = 50$), the average cost saving grow dramatically as the number of samples grows. When the number of features is large ERM+L2 over-performs ERM which suggests that the automatic feature selection is beneficial even with 50 features. In all our

[3] The set of all linear combinations form so-called L-statistics. They are widely used as estimators of quantiles. Note that for the problem of our interests, the optimal solution of the stochastic optimization can be solved analytically if there is no feature, and the optimal solution is a quantile of the unknown net demand distribution.

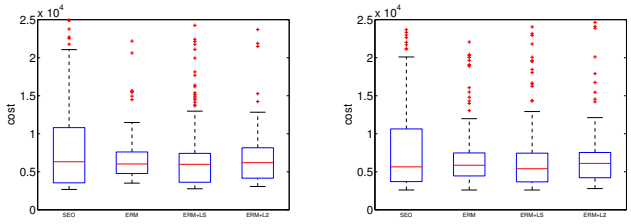
[4] Although this is not a conventional learning problem, in the learning language, we would say the training data set has size n and for each of the $N - n$ experiments the test data set has size 1. Using a larger test set, i.e., deciding the optimal decision for more than one future hours would not make sense in our application as the most recent piece of information has the largest information content regarding the optimal decision.

Table 1. Percentage cost reduction for various number of samples and number of features.

	m=3 ERM	m=3 ERM+LS	m=3 ERM+L2	m=26 ERM	m=26 ERM+LS	m=26 ERM+L2	m=50 ERM	m=50 ERM+LS	m=50 ERM+L2
n=100	29.1	18.9	26.1	22.2	18.7	22.9	8.2	14.3	26.9
n=200	28.4	17.0	24.1	27.7	16.9	24.4	22.7	16.6	28.6
n=300	31.7	16.2	27.0	27.7	15.8	24.7	25.0	15.0	26.8

experiments, ERM+LS does not work well in general for this application, especially comparing to ERM+L2.

As all the cost numbers vary significantly from experiments to experiments, Figure 2 gives the box-plots for two settings, which suggests that all the proposed methods have much smaller spread (variance) in the realized costs and ERM has smallest over the three proposed algorithms.



(a) $n = 100, m = 3$ (b) $n = 100, m = 26$

Figure 2. Box-plots (spreads) of the costs for various methods and scenarios.

4. Conclusion and Future Directions

This project proposes, theoretically analyzes, and empirically tests methods that directly solve stochastic optimization using the data, instead of performing separated estimation and optimization procedure. The application of generator dispatch with renewables is studied in depth to serve as an example of general formulation and methods proposed. Empirical results show that our algorithms lead to great cost savings ($\sim 20\%$) for the energy grid operator.

Future work that develops bounds for regularized version of the algorithm, improves the theoretical understanding on the effect of model mis-specification, investigates how estimators from statistics and operational statistics can serve as features and improve the performance. and tests extension of the algorithms such as kernelized versions are of interest.

A. Proof Sketches

Proof Sketch of Theorem 1. This result relies on a theorem that bounds the excess risk with the sum of 4 times of the Rademacher complexity of the loss function class and $\sqrt{2\log(2/\delta)/n}$ (Bousquet et al., 2004). For the linear hypothesis class with bounded ℓ_2 norm, the Rademacher complexity is bounded by $W^{\max}X^{\max}/\sqrt{n}$. The observation that the loss function is Lipschitz continuous with coefficient $q - c$ translates the Rademacher complexity of the hypothesis class to that of the loss function class and completes the proof. □

Proof Sketch of Theorem 2. As routine algorithmic stability proofs, we have to establish that the algorithm is stable (in fact uniformly stable). One can show (with very tedious algebra) that for our problem and the ERM algorithm, the stability coefficient is $\alpha = D^{\max}(q-c)^2 m/(cn)$. Furthermore, the cost function is uniformly bounded by $K = D^{\max}(q - c)$. Invoking a standard algorithmic stability theorem (Bousquet et al., 2004) gives the bound that

$$|R(w^{\mathrm{ERM}}) - \hat{R}(w^{\mathrm{ERM}})| \leq 2\alpha + (4\alpha n + K)\sqrt{\log(2/\delta)/2n},$$

and completes the proof. □

References

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pp. 169–207. Springer, 2004.

Kao, Y., Roy, B.V., and Yan, X. Directed regression. In *Proc. Advances in Neural Information Processing Systems*, pp. 889–897, 2009.

Lehmann, Erich Leo and Casella, George. *Theory of point estimation*, volume 31. Springer, 1998.

Liyanage, Liwan H. and Shanthikumar, J.George. A practical inventory control policy using operational statistics. *Operations Research Letters*, 33(4):341 – 348, 2005. ISSN 0167-6377.

Qin, J., Su, H., and Rajagopal, R. Storage in risk limiting dispatch: Control and approximation. In *Proc. American Control Conference*, 2013a.

Qin, J., Zhang, B., and Rajagopal, R. Risk limiting dispatch with ramping constraints. In *Proc. IEEE International Conference on Smart Grid Communications (SmartGridComm),*, pp. 791–796, 2013b.