On Min-Max Optimization Over Large Data Sets

Soumyadip Ghosh, Mark S. Squillante
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
{ghoshs, mss}@us.ibm.com

Ebisa D. Wollega
Department of Engineering
Colorado State University-Pueblo
Pueblo, CO 81001
ebisa.wollega@cspueblo.edu

1. INTRODUCTION

We consider a general min-max optimization formulation defined over a sample space $\mathbb{X}$, probability distribution P on $\mathbb{X}$, and parameter space $\Theta \subseteq \mathbb{R}^d$. Define $L_P(\theta) := \mathbb{E}_P[l(\theta, \xi)]$ to be the expectation w.r.t. P of a loss function $l : \Theta \times \mathbb{X} \to \mathbb{R}$ representing the estimation error for a model with parameters $\theta \in \Theta$ over data $\xi \in \mathbb{X}$. Define the expected worst-case loss function $R(\theta) := \mathbb{E}_{P^*(\theta)}[l(\theta, \xi)] = \sup_{P \in \mathcal{P}}\{L_P(\theta)\}$, which maximizes the loss L_P over a set of measures $\mathcal{P}$ of the form

$$\mathcal{P} = \Big\{P \mid D(P, P_b) \leq \rho, \quad \int dP(\xi) = 1, \quad P(\xi) \geq 0\Big\}, \quad (1)$$

where $D(\cdot, \cdot)$ is a measure of distance on the space of probability distributions on $\mathbb{X}$ and the constraints limit the feasible candidates to be within a distance ρ of a base distribution, P_b. Our interest lies in the general ϕ-divergence class of distance measures

$$D_\phi(P, P_b) = \mathbb{E}_{P_b}\left[\phi\left(\frac{dP}{dP_b}\right)\right], \quad (2)$$

where $\phi(u)$ is a non-negative convex function that takes a value of 0 only at $u = 1$. Then, the general min-max formulation consists of solving, for a given $\mathbb{X}$ and $\mathcal{P}$,

$$R(\theta^*_{rob}) = \min_{\theta \in \Theta}\Big\{R(\theta)\Big\} = \min_{\theta \in \Theta}\Big\{\sup_{P \in \mathcal{P}}\{L_P(\theta)\}\Big\}. \quad (3)$$

The formulation in (3) w.r.t. (1) and (2) arises in a wide variety of applications, and has received considerable attention especially recently; see, e.g., [1, 2, 4, 7, 9, 11]. For the χ^2-metric (i.e., $\phi(u) = (u-1)^2$), Namkoong and Duchi [10] show that, for convex and bounded loss functions l with $P_{b,N} = (\frac{1}{N})$ the empirical distribution over a data set of size N, the following result holds with high probability (w.h.p.)

$$\mathbb{E}_{P^*(\theta)}[l(\theta, \xi)] = \mathbb{E}_{P_b}[l(\theta, \xi)] + \sqrt{\frac{\rho \mathrm{Var}_{P_b}(l(\theta, \xi))}{N}}, \quad (4)$$

$\theta \in \Theta$. Similar results have been obtained for other ϕ-divergence metrics such as Kuhlback-Leibler (KL) divergence (i.e., $\phi(u) = u \log u + u - 1$) [7, 4]. An appropriate choice of ρ leads to an optimal solution θ^*_{rob} that has loss performance within $O(\frac{1}{N})$ of the (unknown) true optimal θ^* [10]. This is in contrast to the θ^*_{erm} identified by minimizing $\mathbb{E}_{P_b}[l(\theta, \xi)]$, which leads to a solution with $O(\frac{1}{\sqrt{N}})$ loss performance.

The formulation in (4) over $\theta \in \Theta$ is hard to solve because of the non-convexity of the second term, even if l is strongly convex. Since our general min-max formulation in (3) is convex in θ, and given the (likely) better statistical properties of θ^*_{rob}, an efficient solution procedure is highly desirable. This is especially true when the optimal solution to the inner maximization problem over $\mathcal{P}$ cannot be obtained in closed form, which is the case for the ϕ-divergence constraints in (2). Define the vectors $P := (p_n)$ and $P_{b,N} := (\frac{1}{N})$ of dimension N. We shall focus on the case where P_b is the pmf over a (large) data set of size N, and thus on the loss function and constraint set $\mathcal{P}$ given by $L_{P_b}(\theta) = \frac{1}{N}\sum_{n=1}^N l(\theta, \xi_n)$ and $\mathcal{P} = \{P \mid \sum_{n=1}^N p_n = 1; p_n \geq 0, \forall n; D_\phi(P, P_{b,N}) = \frac{1}{N}\sum_{n=1}^N \phi(N p_n) \leq \rho\}$.

Namkoong and Duchi [10] propose to determine the optimal $P^*(\theta)$ by solving the problem (3) as a deterministic gradient descent problem. They show that, for the χ^2 case, the inner maximization can be reduced to a one-dimensional root-finding problem, solved via bisection search. The key issue is that this bisection search requires an $O(N \log N)$ amount of effort (see Proposition 2) at each iteration, which can be very expensive for large data set sizes N.

2. ALGORITHM AND ANALYSIS

We propose a new primal descent algorithm to solve (3) that is applicable for various ϕ-divergence measures (2). Our solution approach and main results are summarized here; we refer to [6] for all technical details and related work.

For the inner maximization problem, we devise a bisection-search that generalizes the results in [9, 10, 5], showing that our solution applies to many other ϕ-divergence metrics. For the outer minimization problem, instead of operating with the complete data set $M_t = N$ for all iterations t of a gradient descent algorithm, we propose the *stochastic sub-gradient descent* scheme: $\theta_{t+1} = \theta_t - \gamma_t \nabla_\theta \hat{R}_{\mathcal{M}_t}(\theta_t)$ where γ_t is typically called the step-size or learning rate, $\mathcal{M}_t$ is a relatively small subset of the full data set having size $|\mathcal{M}_t| = M_t$, and $\hat{R}_{\mathcal{M}_t}(\cdot)$ is an approximation of $R(\cdot)$ obtained by *uniformly* sampling the subset $\mathcal{M}_t$ from the complete data set of size N. This reduces the computational effort to $O(M_t \log M_t)$ for D_{χ^2}-constrained (3) and $O(M_t)$ for D_{KL}-constraints, where M_t is the size of the support of the pmf in iteration t. (For important reasons related to sampling bias as explained below, we will exploit sampling *without replacement* in our algorithm.)

More precisely, defining $\hat{P} = (\hat{p}_m)$ of dimension M_t, we have the following formulation for our approach

$$\hat{R}_{\mathcal{M}_t}(\theta) = \max_{\hat{P}=(\hat{p}_m)} \sum_{m \in \mathcal{M}_t} \hat{p}_m l(\theta, \xi_m) \quad (5)$$

MAMA 2018 Irvine, California, USA

$$\text{s.t.} \sum_{m\in\mathcal{M}_t} \phi(M_t\hat{p}_m) \le M_t\rho_{M_t},\ \sum_{m\in\mathcal{M}_t} \hat{p}_m = 1,\ \hat{p}_m \ge 0.$$

The cost of solving this problem via bisection search is $O(M_t \log M_t)$. Now suppose $\hat{P}^*(\theta) = (\hat{p}^*_m(\theta))$ is an optimal solution to (5). Then the vector

$$\nabla_\theta \hat{R}_{\mathcal{M}_t}(\theta_t) \quad = \quad \sum_{m\in\mathcal{M}_t} \hat{p}^*_m(\theta_t)\nabla_\theta l(\theta_t,\xi_m) \qquad (6)$$

is a valid sub-gradient for $\hat{R}_{\mathcal{M}_t}(\theta_t)$ and thus we use it in our stochastic sub-gradient descent. We apply this to two specific ϕ-divergences, noting that the optimal value $\hat{R}^*_{\mathcal{M}_t}(\theta)$ for many other ϕ-divergences can be obtained similarly.

PROPOSITION 1. *The optimal solution $\hat{P}^*$ to the problem* (5) *with a KL-divergence constraint ($\phi(u) = u\log u - (u-1)$) is given by: (a)* Case $\alpha^* = 0$*: $\hat{p}^*_m = \frac{1}{M'_t}$, where $M'_t = |\mathcal{M}'_t|$, $\mathcal{M}'_t = \{m \in \mathcal{M}_t | z_m = \max_u z_u\}$; (b)* Case $D_{KL}(\hat{P}^*, P_b) = \rho_{M_t}$*: $\hat{p}^*_m = \frac{e^{z_m\beta^*}}{\sum_j e^{z_j\beta^*}}$, where β^* solves $\beta\kappa'(\beta) - \kappa(\beta) = \rho_{M_t}$ and $\kappa(\beta) = \sum_j e^{z_j\beta}/M_t$. The computational complexity is $O(M_t\log(1/\epsilon))$, where ϵ is the desired accuracy.*

PROPOSITION 2. *An optimal solution to the problem* (5) *with a χ^2-divergence constraint ($\phi(u) = (u-1)^2$) is given by: (a)* Case $\alpha^* = 0$*: $\hat{p}^*_m = \frac{1}{M'_t}$, where $M'_t = |\mathcal{M}'_t|$, $\mathcal{M}'_t = \{m \in \mathcal{M}_t | z_m = \max_u z_u\}$; (b)* Case $D_{KL}(\hat{P}^*, P_b) = \rho_{M_t}$*: $\hat{p}^*_m = \begin{cases} \frac{z_m - z_{\min} - \lambda^*}{2\alpha^* M_t} + \frac{1}{M_t}, & z_m \ge z_{\min} + \lambda^* - 2\alpha, \\ 0, & z_m < z_{\min} + \lambda^* - 2\alpha, \end{cases}$ where $z_{\min} = \min_m z_m$ and λ^*, α^* jointly solve: $\sum_m (z_m - z_{\min})\ \mathbb{I}\{\hat{p}^*_m > 0\} - 2M_t\alpha^* = (\lambda^* - 2\alpha^*)\big(\sum_m \mathbb{I}\{\hat{p}^*_m > 0\}\big)$ and $\sum_m (z_m - z_{\min} - \lambda^*)^2\ \mathbb{I}\{\hat{p}^*_m > 0\} = 4\rho_{M_t}M_t$. Furthermore, the computational complexity to obtain the primal-dual optimal solutions $\hat{P}^*, \alpha^*, \lambda^*$ is $O((M_t - \log\epsilon)\log M_t)$, where ϵ is the desired precision.*

Next, the quality of the approximation of $\hat{R}^*_{\mathcal{M}_t}$, or more particularly that of its sub-gradient, is a primary concern with this approach. We show in Theorem 3 that the bias induced by the subsampling of the full support is of order $O((1/M_t - 1/N)^{1-\delta})$. Since this estimator is biased and the only control on it is via M_t, our method necessarily grows $M_t \nearrow N$ as t increases. The result specifically depends on the $\mathcal{M}_t$ being sampled without replacement, and Remark 1 below explains why sampling with replacement makes the method computationally burdensome.

We restrict our attention to ϕ-divergences that satisfy, for a small $\zeta > 0$, the continuity condition $|\phi(u(1+\zeta)) - \phi(u)| \le \kappa_1(\zeta)\phi(u) + \kappa_2(\zeta)$, where $\kappa_1(\zeta)$ and $\kappa_2(\zeta)$ are both $O(\zeta)$. This continuity condition can be verified for many common ϕ-divergence measures of interest including the χ^2 and KL-divergence metrics. Let $\mathbb{E}_{\mathcal{M}_t}$ and $\mathbb{P}_{\mathcal{M}_t}$ be expectations and probabilities w.r.t. the uniform sampling without-replacement producing the random set $\mathcal{M}_t$.

THEOREM 3. *Suppose the optimal solution P^* to (3) is unique and $\rho \ll 1$ in* (1)*. Assume the ϕ-divergence satisfies the continuity condition and define the D_ϕ-constraint target in* (5) *to be $\rho_{M_t} = \rho + \eta_{M_t}$, where $\eta_{M_t} = c\big(\frac{1}{M_t} - \frac{1}{N}\big)^{(1-\delta)/2}$ for constant $c > 0$ and small constant $\delta > 0$. Then, for all $M_t \ge M_0$ with M_0 sufficiently large, we have that the sub-gradient $\nabla_\theta\hat{R}_{\mathcal{M}_t}(\theta)$ and full-gradient $\nabla_\theta R(\theta)$ satisfy $\|\mathbb{E}_{\mathcal{M}_t}[\ \nabla_\theta\hat{R}_{\mathcal{M}_t}(\theta)] - \nabla_\theta R(\theta)\|_2^2 \to C\eta^2_{M_t}$ as $M_t \to N$, where C is a finite constant.*

We provide here a sketch of the proof. First construct $\tilde{P}_{\mathcal{M}_t} = (\tilde{p}_1, \ldots, \tilde{p}_{M_t})$, a restriction of the (unique) optimal solution P^* of the full-data problem (3) onto the (random) subset $\mathcal{M}_t$ of support points used in the restricted problem (5), where $\tilde{p}_m = \frac{p^*_m}{\sum_{j\in\mathcal{M}_t} p^*_j}$, $\forall m \in \mathcal{M}_t$. The condition $\rho \ll 1$ ensures that, w.h.p., the summation in the denominator is greater than zero for a sufficiently large M_t. We then show that, w.h.p. (under the $\mathcal{M}_t$-sampling measure), the pmf $\tilde{P}_{\mathcal{M}_t}$ is a feasible solution to (5) when ρ_{M_t} is inflated as assumed. Next, we establish that $\mathbb{E}_{\mathcal{M}_t}[|z^T(\tilde{P}_{\mathcal{M}_t} - P^*)|]$ is of the order $O(\eta^{2/(1-\delta)}_{M_t})$, where z^T denotes the transpose of vector z. Since $\tilde{P}_{\mathcal{M}_t}$ is a feasible solution to (5) for sufficiently large $\mathcal{M}_t$ and a slightly larger ρ, an appeal to the fundamental theorem of calculus yields the desired result.

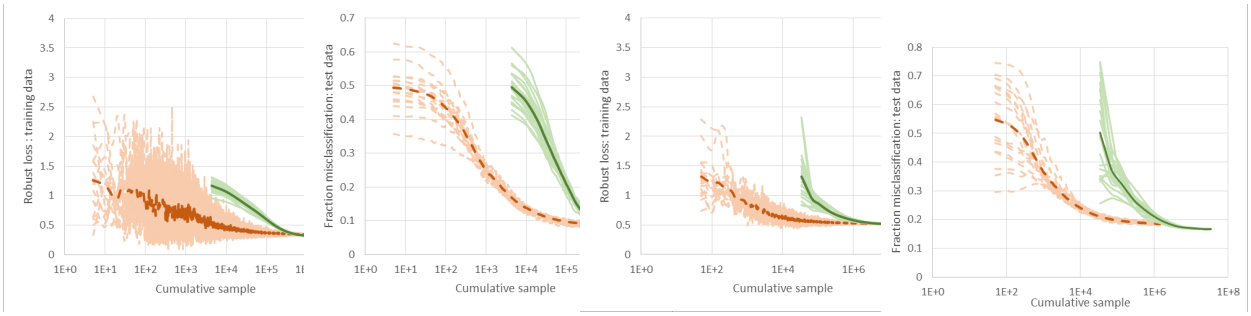
REMARK 1. *The squared bias in Theorem 3 is more accurately stated as $O(\eta^2_{|\mathcal{M}_t|})$, where $|\mathcal{M}_t|$ is the number of support points used in* (5)*. We require sampling without replacement because M_t samples* with replacement *only produces a set $\mathcal{M}_t$ such that $|\mathcal{M}_t| = O(\log M_t)$. The resulting slow drop in bias makes the method inefficient in terms of the computational effort expended.*

Lastly, we look at sample size growth rules where the maximum size N is hit after a (large but) finite number of iterations, addressing the question of balancing the added computational burden of each iteration against the expected reduction in optimality gap. We assume: (a) For each $\xi_n,\ n = 1, \ldots, N$, the loss functions $l(\theta, \xi_n)$ are c-strongly convex and their gradients $\nabla_\theta l(\theta, \xi_n)$ are L-Lipschitz; Additionally, the Hessian $\nabla^2_\theta l(\theta, \xi_n)$ exists. (b) The robust loss function $R(\theta)$ has a unique minimizer θ_{rob} that satisfies (3); (c) The estimator $\nabla_\theta R(\theta)$ obeys a bound $\mathbb{E}_{\mathcal{M}_t}\big[\|\nabla_\theta\hat{R}_{\mathcal{M}_t}(\theta) - \mathbb{E}_{\mathcal{M}_t}[\ \nabla_\theta\hat{R}_{\mathcal{M}_t}(\theta)]\ \|_2^2\big] \le C'\eta^{2/(1-\delta)}_{M_t}$.

Then Proposition 4 shows that the properties in assumption (a) translate over to the robust performance metric $R(\theta)$ as defined in (3). We can relax the assumption to have sample-dependent constants $c(\xi)$, $L(\xi)$. Since the number of samples is finite, $\bar{L} = \max_\xi L(\xi)$ and $\underline{c} = \min_\xi c(\xi)$ are sample-independent values that can be used in place of c, L in Proposition 4 to obtain the same properties.

PROPOSITION 4. *Under the above assumptions, the function $R(\theta) = \max_{P\in\mathcal{P}} L_P(\theta)$ is c-strongly convex, and its gradient $\nabla_\theta R(\theta)$ is L-Lipschitz.*

Solving (3) with a deterministic line search algorithm provides a linear reduction in error but with a per-iteration level of effort on the order of $O(N)$. Alternatively, the standard prescription from stochastic gradient descent (SGD) algorithms is that the sample size be maintained at a constant $M_t = M \ll N$ throughout the iterations. This, however, would lead to biased sampling in the iterates given Theorem 3, which provides only $M_t \nearrow N$ as a control. Fixed bias violates a basic requirement for SGD that the gradient estimator $\nabla_\theta\hat{R}_{\mathcal{M}_t}(\theta) = \Theta(\nabla_\theta R(\theta))$ (see, e.g., 4.3 in [3]). Hence, convergence of our sub-gradient descent scheme cannot be guaranteed when $M_t = M,\ \forall t$. In the following theorem, we show that a growing M_t ensures convergence. Additionally, our choice of geometrically increasing M_t and fixed step sizes γ is efficient w.r.t. the total computational budget W_t that is expended up until iterate t, which is the sum of the amount of individual work w_t in



(a) HIV-1: $\rho = 0.1$, train (b) HIV-1: $\rho = 0.1$, test (c) Adult: $\rho = 0.1$, train (d) Adult: $\rho = 0.1$, test

Figure 1: HIV-1 and Adult comparisons of sub-gradient (orange, dashed) and full-gradient (green, solid) algorithms.

each iterate. Define the ratio $\nu_t := M_t/M_{t+1}$ as the *growth factor* of the sequence $\{M_t\}$, and the expected optimality gap $O_{t+1} := \mathbb{E}_{\mathcal{M}_t}[R(\theta_{t+1})] - R(\theta_{\text{rob}})$ which drops as $M_t \nearrow N$.

THEOREM 5. *Suppose the constant step-size $\gamma_t = \gamma$ satisfies $\gamma \leq \min\{\frac{1}{4L}, 4c\}$, $\forall t$. Let $r = 1 - \frac{\gamma}{4c}$. We then have: (a) If M_t grows geometrically with parameter $\nu < r$, then for $t \leq t_{\max}$, $O_{t+1} = O(r^t)$. Further, if we use D_{KL}-constraints, then $O_{t+1} = O(W_t^{-1}(r/\nu)^t)$, and if D_{χ^2}-constraints are used then $O_{t+1} = O(W_t^1(r/\nu)^t/t)$; (b) If M_t grows sub-geometrically, then $O_{t+1} = O(w_t^{-1})$.*

Theorem 5 establishes that any sub-geometric rate of growth will lead to sub-optimal reduction in the optimality gap w.r.t. the total computational effort. Intuitively, this can be understood to happen because the stochastic error drops to zero much slower than the deterministic error that can be attained for strongly convex optimization objectives, and thus the stochastic error dominates. This yields that sub-geometric rates are suboptimal in the sense that $W_t O_{t+1} \to \infty$ as $t \nearrow$, indicating that the error O_t is unable to drop fast enough compared to the rate at which W_t grows.

Geometrically increasing the sampling will, on the other hand, attain a balance between the rate of convergence of the stochastic error and the deterministic improvement possible for strongly convex functions, thus attaining a better balance between the optimality gap and the level of computational effort. The fastest convergence is attained when $\nu = r$, eliminating the $(r/\nu)^t$ inflation factor. However, r depends on c and L through γ, so it is hard to obtain in practice.

3. EXPERIMENTAL RESULTS

Numerous experiments were conducted to empirically evaluate our new SGD algorithm in comparison with the full-gradient algorithm based on two data sets from [8]: HIV-1 Protease Cleavage; and Adult Income. Additional details on both data sets for our experiments is provided in [6].

Each data set was split by randomly selecting 25% of the data selected for testing, with the remaining data used for training. Empirical results from twenty experimental runs under our sub-gradient algorithm and under the full-gradient algorithm are presented in Figure 1(a) and (b) for the HIV-1 data set and in Figure 1(c) and (d) for the Adult Income data set; the bold dashed and solid lines respectively illustrate the average of the twenty runs. The constraint parameter ρ was set to 0.1, with the corresponding figures for different values of ρ provided in [6]. The leftmost plots compare the robust loss performance objective of the two algorithms based on the training testing data, and the rightmost plots compare the fractional misclassification performance of the two algorithms based on the testing data. We observe from these experiments that our proposed SGD algorithm outperforms the full-gradient method for each value of ρ considered, with the best results obtained for $\rho = 0.1$. We further note that our method can become somewhat unstable when the ρ value becomes sufficiently large, which is consistent with our above results prescribing $\rho < 1$.

4. REFERENCES

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

[2] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein Profile Inference and Applications to Machine Learning. *ArXiv e-prints*, Oct. 2016.

[3] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *ArXiv e-prints*, June 2016.

[4] J. Duchi, P. Glynn, and H. Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *ArXiv e-prints*, Oct. 2016.

[5] S. Ghosh and H. Lam. Robust analysis in stochastic simulation: Computation and performance guarantees. *Operations Research*. To appear, 2018.

[6] S. Ghosh, M. S. Squillante, and E. D. Wollega. Efficient Stochastic Gradient Descent for Distributionally Robust Learning. *ArXiv e-prints*, May 2018.

[7] H. Lam. Robust sensitivity analysis for stochastic systems. *Math. Op. Res.*, 41(4):1248–1275, 2016.

[8] M. Lichman. UCI machine learning repository, 2013.

[9] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, 29:2208–2216, 2016.

[10] H. Namkoong and J. C. Duchi. Variance-based regularization with convex objectives. In *NIPS*, 30:2971–2980, 2017.

[11] A. Sinha, H. Namkoong, and J. Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. In *ICLR*, 2018.

[12] S. S. Wilks. *Mathematical Statistics*. Wiley, 1962.