KYBERNETIKA — VOLUME 57 (2021), NUMBER 1, PAGES 38–45

A NOTE ON THE CONVERGENCE RATE IN REGULARIZED STOCHASTIC PROGRAMMING

Evgueni Gordienko and Yury Gryazin

We deal with a stochastic programming problem that can be inconsistent. To overcome the inconsistency we apply Tikhonov's regularization technique, and, using recent results on the convergence rate of empirical measures in Wasserstein metric, we treat the following two related problems:

1. A choice of regularization parameters that guarantees the convergence of the minimization procedure.
2. Estimation of the rate of convergence in probability.

Considering both light and heavy tail distributions and Lipschitz objective functions (which can be unbounded), we obtain the power bounds for the convergence rate.

Keywords: stochastic programming problem, Tikhonov's regularization, Lipschitz conditions, Kantorovich metric, convergence rate

Classification: 90C15

1. PROBLEM SETTING

We consider a stochastic optimization problem associated with, so-called, empirical risk minimization in the learning theory (see, for instance [2, 3, 11, 14, 15]). We assume that all random vectors and variables, which appear in this paper, are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Let $\bar{X} = (X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random vector with values $X \in \mathcal{X} \subset \mathbb{R}^k$ and $Y \in \mathcal{Y} \subset \mathbb{R}^m$, where $\mathcal{X}$ and $\mathcal{Y}$ are given Borel sets.

Let $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be a given measurable function (*loss function* in learning, that is related to *objective function* in "classical" stochastic optimization), and $F = \{f : \mathcal{X} \to \mathcal{Y}\}$ be a given class of measurable functions. We suppose that for each $f \in F$,

$$R(f) := E\mathcal{L}(Y, f(X)) < \infty \tag{1}$$

and there exists $f_* \in F$ such that

$$R^* := \inf_{f \in F} R(f) = R(f_*). \tag{2}$$

DOI: 10.14736/kyb-2021-1-0038

One of typical problems in stochastic optimization is estimating R^* in (2) in the situation where the distribution $\boldsymbol{P} := \boldsymbol{P}_{(X,Y)}$ of (X, Y) is unknown, but a sample

$$\{\bar{X}_1 = (X_1, Y_1), \ldots, \bar{X}_n = (X_n, Y_n)\} \tag{3}$$

of i.i.d. random vectors with a common law $\boldsymbol{P}$ is available.

The standard approach to tackle the problem is the following (see, for example, [2, 3, 7, 8, 9, 10, 12, 14]).

Let $\boldsymbol{P_n} := \frac{1}{n} \sum\limits_{i=1}^{n} \delta_{(X_i,Y_i)}$ be the corresponding empirical distribution,

$$R_n(f) := E_{\boldsymbol{P}_n} \mathcal{L}(Y, f(X)) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(Y_i, f(X_i)), \quad f \in F, \tag{4}$$

and

$$R_n^* := \inf_{f \in F} R_n(f), \quad n = 1, 2, \ldots \tag{5}$$

There is a vast literature about conditions (for example, boundedness, compactness, continuity, or moment conditions) that make the problem *consistent*. That is

$$R_n^* \to R^* \ \text{ as } \ n \to \infty \tag{6}$$

(either in probability, or with probability one).

Moreover, considering a "classical" version of the stochastic programming problem (1) – (5) with an objective function $\mathbb{L} : \mathbb{R}^s \times \mathbb{R}^k \to \mathbb{R}$ and

$$R^* := \inf_{x \in \mathfrak{X} \subset \mathbb{R}^s} E\mathbb{L}(x, \xi), \tag{7}$$

where $x \in \mathbb{R}^s$ is a "decision vector", and ξ is a $k-$dimensional random vector with an unknown distribution, one can find many papers (possibly starting in 1978 by paper [7]) on estimations of the rate of convergence of R_n^* to R^* (in probability, see, for instance [2, 8, 9, 10, 12, 14]).

Our goal also is the estimation of the convergence rate (in probability), but in situations where the consistency (6) can fail to hold, and a *regularization* is needed to make the problem well-posed.

Note that the technique and the results obtained in this paper can (after appropriate modifications) be applied to the "classical" stochastic programming, where R^* in (7) is estimated by using empirical distributions.

2. ASSUMPTIONS AND EXAMPLE

The Lipschitz (or Hölder) conditions on a loss function (or on a objective function) are common in stochastic programming ([5, 7, 9, 10, 12]) and the learning theory ([2, 3, 11, 14, 15]). At the same time, the Lipschitz conditions on functions f from F in (1) and (2) are also in use (for example, considering linear functions or splines of first order, [2, 3, 5, 11, 14, 15].

Let $|\cdot|_k$ and $|\cdot|_m$ be Euclidian norms in $\mathbb{R}^k$ and $\mathbb{R}^m$, respectively. A norm on $\mathbb{R}^d = \mathbb{R}^k \times \mathbb{R}^m$ is defined by $|\cdot| := |\cdot|_k + |\cdot|_m$.

Let

$$\mathcal{F} := \{f : \mathcal{X} \to \mathcal{Y} \text{ such that } \|f\|_L := \sup_{x \neq x'} \frac{|f(x) - f(x')|_m}{|x - x'|_k} < \infty\} \tag{8}$$

be the class of all maps from $\mathcal{X}$ to $\mathcal{Y}$ with a finite *Lipschitz norm* $\|\cdot\|_L$.

Assumption 1. There is a constant $L < \infty$ such that for any $y, \tilde{y}, y'$ and $\tilde{y}' \in \mathcal{Y}$

$$|\mathcal{L}(y, \tilde{y}) - \mathcal{L}(y', \tilde{y}')| \leq L(|y - y'|_m + |\tilde{y} - \tilde{y}'|_m). \tag{9}$$

Assumption 2. F is a subset of $\mathcal{F}$.

Assumption 3. For some $\beta \geq 3$

$$E|X|_k^\beta < \infty, \quad E|Y|_m^\beta < \infty. \tag{10}$$

Example 1. Let $k = m = 1$, $\mathcal{X} = [1, 3]$, X be a random variable distributed uniformly on $[1, 3]$, and for every Borel set $B \subset \mathbb{R}$,

$$P(Y \in B | X = x) = \begin{cases} \mathbb{P}(\eta_x \in B) & x \in [1, 2], \\ \mathbb{P}(\eta_{2x-2} \in B) & x \in (2, 3]. \end{cases}$$

Here $\eta_x \sim \mathit{Norm}(x, 1)$, $\eta_{2x-2} \sim \mathit{Norm}(2x - 2, 1)$, (that is, the law of Y conditionally to X is a Gaussian centered either in X or in $2X - 2$.)

Let $\mathcal{L} = |y - f(x)|$ and $F := \{f : [1, 3] \to \mathbb{R} : \|f\|_L < \infty\}$. Then Assumptions 1-3 are satisfied, and in (1)

$$R(f) = \frac{1}{2} \int_1^3 E\left\{|Y - f(X)| \,|\, X = x\right\} \mathrm{d}x$$

$$= \frac{1}{2} \int_1^2 E\,|\eta_x - f(x)|\,\mathrm{d}x + \frac{1}{2} \int_2^3 E\,|\eta_{2x-2} - f(x)|\,\mathrm{d}x. \tag{11}$$

Also,

$$\inf_{f \in F} E|\eta_x - f(x)| = \inf_{z \in \mathbb{R}} E|\eta_x - z| = \mathit{median}(\eta_x) = x, \qquad x \in [1, 2]$$

and

$$\inf_{f \in F} E|\eta_{2x-2} - f(x)| = \mathit{median}(\eta_{2x-2}) = 2x - 2, \qquad x \in (2, 3].$$

Consequently, $\inf\limits_{f \in F} R(f)$ is attained at the function

$$f_*(x) = \begin{cases} x & , \ x \in [1, 2], \\ 2x - 2 & , \ x \in (2, 3]. \end{cases}$$

From (11) we see that ($\eta \sim \mathrm{Norm}(0, 1)$) $R^* = R(f_*) = E|\eta| = \sqrt{\frac{2}{\pi}}$.

The solution of the problem in (4), (5) provides a quite different result. Choosing in (4) the piecewise linear function f_n^*, that connects the points Y_i, $i = 1, \ldots, n$ gives $f(X_i) = Y_i$ for all i. Therefore, in (5) $R_n^* = R_n(f_n^*) = 0$ for all $n = 1, 2, \ldots$. Note that with probability one $\|f_n^*\|_L \to \infty$ as $n \to \infty$.

3. REGULARIZATION

The matter of fact we observed in Example 1 is known as *overfitting* (in learning theory and other areas of stochastic optimization). To cope with this phenomenon, the different methods of *regularization* are applied. (See, for instance, [3, 5, 11, 15]).

In our rather general setting, when we do not consider particular minimization algorithms and treat the Lipschitz case, Tikkonov's regularization (see, for instance, [6, 11, 13, 15])

$$\tilde{R}_n(f) := R_n(f) + \alpha_n V(f) \tag{12}$$

is adequate. In (12) R_n was defined in (4), and $V(f) \equiv V(\|f\|_L)$ is the regularization term, that penalises "too large slopes" $\|f\|_L$ of $f \in F$.

In the above setting the following two problems seem to be new:

(1) A suitable choise of the *regularization parameters* α_n in (12), which takes into account the error of approximation of $\boldsymbol{P}$ by $\boldsymbol{P_n}$, and guaranties

$$\inf_{f\in F} \tilde{R}_n(f) \xrightarrow{\mathbb{P}} R^* \text{ as } n \to \infty. \tag{13}$$

(2) Estimation of the rate of convergence in (13).

4. TIKHONOV'S REGULARIZATION AND THE CONVERGENCE RATE

Let (L is defined in (9))

$$V(f) := L \max\{1,\ \|f\|_L\}, \quad f \in F, \tag{14}$$

and, given $\{\alpha_n\} \subset (0, \infty)$, for $n = 1, 2, \ldots$

$$G_{n,\alpha_n}(f) := R_n(f) + \alpha_n V(f), \quad f \in F, \tag{15}$$

with $R_n(f)$ being defined in (4).

First, we need to choose a suitable measure of an error of the approximation of the unknown $\boldsymbol{P}$ by means of $\boldsymbol{P_n}$.

For probabilities measures μ and ν on $\mathcal{W} := \mathcal{X} \times \mathcal{Y}$ (with finite first moments) the Kantorovich (or Wasserstein of order 1) distance is

$$\varkappa(\mu, \nu) := \sup_{\phi\in\Phi} |\int_{\mathcal{W}} \phi(x)\mu(\mathrm{d}x) - \int_{\mathcal{W}} \phi(x)\nu(\mathrm{d}x)|, \tag{16}$$

where $\Phi := \{\phi : \mathcal{W} \to \mathbb{R} : \|\phi\|_L \le 1\}$.

The below proposition is a particular case of Theorem 1 in [4].

Proposition 1. Under Assumption 3 there is a constant $c_\beta < \infty$ such that,

$$E\varkappa(\boldsymbol{P}, \boldsymbol{P_n}) \le c_\beta\, \gamma_n, \qquad n = 1, 2, \ldots, \tag{17}$$

where for $n = 1, 2, \ldots$

$$\gamma_n = \begin{cases} n^{-\frac{1}{2}} \log(1+n) & \text{if} \quad d := k + m = 2, \\ n^{-\frac{1}{d}} & \text{if} \quad d \ge 3. \end{cases} \tag{18}$$

The rate of convergence in (17) is the best possible, and c_β can be estimated in terms of β and d (see [4]).

Given some known bound $c > c_\beta$ and fixing

$$\alpha \in \left(0, \frac{1}{2} - \frac{1}{\beta}\right), \tag{19}$$

we define:

$$\alpha_n := c\,\gamma_n + n^{-\alpha}, \qquad n = 1, 2, \dots \tag{20}$$

In place of the generally ill-posed problem (5), we consider the following (*regularized*) problem of minimization of "empirical risk" .

Let $\{f_n\} \subset F$ be any sequence such that (see (15))

$$G_{n,\alpha_n}(f_n) \le \inf_{f\in F} G_{n,\alpha_n}(f) + \alpha_n\, h_n, \qquad n = 1, 2, \dots \tag{21}$$

In (21) $\{h_n\}$ is a sequence of nonnegative random variables. The term $\alpha_n h_n$ represents a possible error of minimization of G_{n,α_n}. We assume

$$\sup_{n\ge 1} Eh_n = h_* < \infty. \tag{22}$$

Recall that $R(f)$ and R^* were introduced in (1) and (2).

Theorem 1. Let Assumtions 1-3 hold and $\{f_n\}$ be defined in (21). Then the sequence $R(f_n),\ n \ge 1$ converges in probability to $R^* = \inf\limits_{f\in F} R(f)$. Moreover, for every $\varepsilon > 0$, $n \ge 1$

$$\mathbb{P}(R(f_n) - R^* > \varepsilon) \le \frac{1}{\varepsilon}[(ch_* + q(c+1))\gamma_n + (h_* + q)n^{-\alpha}] + \bar{c}\, n^{-\lambda}\{1 + \lambda^{\frac{\beta}{2}}[\log n]^{\frac{\beta}{2}}\}, \tag{23}$$

where

$$q = L\max\{1,\ \|f_*\|_L\}, \tag{24}$$

$$\lambda = \frac{\beta}{2} - \alpha\beta - 1 > 0, \tag{25}$$

c comes from (20) and $\bar{c} := 2^{\frac{5\beta}{2}-1}\{E|X|_k^\beta + E|Y|_m^\beta\}$. The sequence $\{\gamma_n\}$ was defined in (18).

Remark 1. Consider, for instance, the case $k = m = 1$, and suppose that Assumption 3 holds for all $\beta > 0$. Then we can single out a "large enough" β to be able to choose in (19), (20) $\alpha = \frac{1}{2} - \eta$ with a small positive η. In view of (18) $\gamma_n = o(n^{-\alpha})$ as $n \to \infty$, and in (25) $\lambda = \beta\eta - 1$ would be greater than $\frac{1}{2}$. Consequently, the right-hand side of (23) could be $\frac{1}{\varepsilon}O(n^{-\alpha})$ with $\alpha \approx \frac{1}{2}$.

Remark 2. In the situation described in Remark 1 (with $\alpha = \frac{1}{2} - \eta$) take in (23) $\varepsilon = tn^{-\frac{1}{2}+2\eta}$. Then $\mathbb{P}\left((n^{\frac{1}{2}-2\eta})[R(f_n) - R^*] > t\right) \to 0$ as $n \to \infty$.

This assertion resembles the estimations of the convergence rate obtained in the works by V. Kaňková (particularly, Theorem 4.2 in [9]).

The above regularization technique can be applied to the "standard" stochastic optimization problems, where as in (7) one is looking for $\inf_{x\in\mathfrak{X}\subset\mathbb{R}^s} E\mathcal{L}(x,\xi)$, where the set $\mathfrak{X}$ can be noncompact and the objective function $\mathcal{L}:\mathbb{R}^s\times\mathbb{R}^k\to\mathbb{R}$ might be unbounded, however, for each $x\in\mathfrak{X}$ the function $\mathcal{L}(x,\cdot):\mathbb{R}^k\to\mathbb{R}$ is Lipschitzian.

P r o o f . [Proof of the Theorem 1] Given any $f\in F$, from (9) and (8) we obtain

$$\begin{aligned}&|\mathcal{L}(y,f(x))-\mathcal{L}(y',f(x'))|\le L\,(|y-y'|_m+\|f\|_L|x-x'|_k)\\&\le L\max\{1,\ \|f\|_L\}\ |(y,x)-(y',x')|=V(f)\ |(y,x)-(y',x')|.\end{aligned}$$

Hence $\mathcal{L}(\,\cdot\,,f(\cdot))$ is Lipschitz on $\mathcal{W}$. Letting $\delta_n:=\varkappa(\boldsymbol{P},\boldsymbol{P_n})$ by (1) and (16),

$$|R(f)-R_n(f)|\le\delta_n\,V(f),\qquad f\in F. \tag{26}$$

From (21) we infer the following inequalities (see (15)),

$$\begin{aligned}G_{n,\alpha_n}(f_n)&=R_n(f_n)+\alpha_nV(f_n)\le\inf_{f\in F}[R_n(f)+\alpha_n\,V(f)]+\alpha_nh_n\\&\le R_n(f_*)+\alpha_nV(f_*)+\alpha_nh_n.\end{aligned} \tag{27}$$

By (26) and (27)

$$\begin{aligned}&R(f_n)-\delta_nV(f_n)+\alpha_nV(f_n)\le R_n(f_n)+\alpha_nV(f_n)\\&\le R_n(f_*)+\alpha_nV(f_*)+\alpha_nh_n\le R(f_*)+\delta_nV(f_*)+\alpha_nV(f_*)+\alpha_nh_n\\&=R^*+(\alpha_n+\delta_n)V(f_*)+\alpha_nh_n.\end{aligned}$$

From the last inequalities we obtain

$$0\le R(f_n)-R^*\le\alpha_nh_n+(\alpha_n+\delta_n)q+(\delta_n-\alpha_n)V(f_n), \tag{28}$$

where $q:=V(f_*)\equiv L\max\{1,\ \|f_*\|_L\}$.

Let $\varepsilon>0$ and ξ_n denote the sum of the first two terms on the right-hand side of (28). Then $\{\xi_n\le\varepsilon,(\delta_n-\alpha_n)V(f_n)\le0\}\subset\{\xi_n+(\delta_n-\alpha_n)V(f_n)\le\varepsilon\}$, or, passing to the complements, $\{\xi_n+(\delta_n-\alpha_n)V(f_n)>\varepsilon\}\subset\{\xi_n>\varepsilon\}\cup\{(\delta_n-\alpha_n)V(f_n)>0\}$.

Since $V(f_n)$ is strictly positive (see(14)), $\{(\delta_n-\alpha_n)V(f_n)>0\}=\{\delta_n>\alpha_n\}$. Therefore,

$$\mathbb{P}(R(f_n)-R^*>\varepsilon)\le\mathbb{P}(\xi_n>\varepsilon)+\mathbb{P}(\delta_n>\alpha_n). \tag{29}$$

By Markov's inequality, (17), (20) and (22)

$$\begin{aligned}\mathbb{P}(\xi_n>\varepsilon)\le\frac{1}{\varepsilon}[\alpha_nh_*+q(\alpha_n+c\gamma_n)]\le\\\frac{1}{\varepsilon}[(c\,h_*+qc+q)\gamma_n+(h_*+q)n^{-\alpha}].\end{aligned} \tag{30}$$

To bound the second summand on the right-hand side of (29), we apply the following concentration inequality given in Proposition A.2 in [1] which holds under Assumption 3:

$$\mathbb{P}(\varkappa(\boldsymbol{P},\boldsymbol{P_n})\ge E\varkappa(\boldsymbol{P},\boldsymbol{P_n})+t)\le c_1n^{1-\frac{\beta}{2}}t^{-\beta}\left(1+[\log(n^{\frac{\beta}{2}-1}t^{\beta})]^{\frac{\beta}{2}}\right). \tag{31}$$

As it is shown in [1], $c_1 = 2^{\frac{3\beta}{2}} E|(X,Y)|^\beta$. But $E|(X,Y)|^\beta = E[|X|_k + |Y|_m]^\beta \le 2^{\beta-1}\{E|X|_k^\beta + E|Y|_m^\beta\}$. Thus, $c_1 \le 2^{\frac{5\beta}{2}-1}\{E|X|_k^\beta + E|Y|_m^\beta\}$.

We have (see (20))

$$\mathbb{P}(\delta_n > \alpha_n) = \mathbb{P}(\varkappa(\boldsymbol{P}, \boldsymbol{P_n}) > c\gamma_n + n^{-\alpha}) \le \mathbb{P}(\varkappa(\boldsymbol{P}, \boldsymbol{P_n}) > E\varkappa(\boldsymbol{P}, \boldsymbol{P_n}) + n^{-\alpha}), \quad (32)$$

where the last inequality is due to (17) and the above selection of the constant c.

Applying (31) to (32) with $t = n^{-\alpha}$, we get

$$\mathbb{P}(\delta_n > \alpha_n) \le c_1 n^{-(\frac{\beta}{2}-\alpha\beta-1)}\{1 + [\log(n^{\frac{\beta}{2}-\alpha\beta-1})]^{\frac{\beta}{2}}\}.$$

Under condition (19) $\lambda := \frac{\beta}{2} - \alpha\beta - 1 > 0$, and

$$\mathbb{P}(\delta_n > \alpha_n) \le c_1 n^{-\lambda}\{1 + \lambda^{\frac{\beta}{2}}[\log n]^{\frac{\beta}{2}}\}. \quad (33)$$

Joining the inequalities (29), (30) and (33) we obtain the desired inequality (23). □

Example 2. (*Continuation of Example 1)* For the problem described in Example 1 the inequalities (10) hold true for every $\beta > 3$. So, as it was explained in Remark 1, the parameter α in (20) can be selected close to $\frac{1}{2}$, and the right-hand side of inequality (25) can be made to be $\frac{1}{\varepsilon}O(n^{-\frac{1}{2}+\eta})$, with η near zero.

ACKNOWLEDGMENT

We are grateful to the referees for their valuable suggestions and remarks.

(Received December 30, 2019)

REFERENCES

[1] E. Boissard and T. Le Gouic: On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. Annales de l'Institut Henry Poincaré, Probabilités et Statistiques *50* (2014), 539–563. DOI:10.1214/12-AIHP517

[2] L. Devroye, L. Györfi, and G. Lugosi: A Probabilistic Theory of Pattern Recognition. Springer, New York 1996. DOI:10.1007/978-1-4612-0711-5

[3] T. Evgeniou, T. Poggio, M. Pontil, and A. Verri: Regularization and statistical learning theory for data analysis. Comput. Statist. Data Anal. *38* (2002), 421–432. DOI:10.1016/S0167-9473(01)00069-X

[4] N. Fournier and A. Guillin: On the rate of convergence in Wasserstein distance of the empirical measure. Probab. Theory Related Fields *162* (2015), 707–738. DOI:10.1007/s00440-014-0583-7

[5] E. Gordienko: A remark on stability in prediction and filtering problems. Izv. Akad. Nank SSR Tekhn. Kibernet. *3* (1978), 202–205.

[6] Y. A. Gryazin, M. V. Klibanov, and T. R. Lucas: Numerical solution of a subsurface imaging inverse problem. SIAM J. Appl. Math. *62* (2001), 664–683. DOI:10.1137/S0036139900377366

[7] V. Kaňková: An approximative solution of stochastic optimization problem. In: Trans. 8th Prague Conf. Academia, Prague 1978, pp. 349–353. DOI:10.1007/978-94-009-9857-5_33

[8] V. Kaňková: Empirical estimates in stochastic programming via distribution tails. Kybernetika *46* (2010), 459–471.

[9] V. Kaňková and M. Houda: Thin and heavy tails in stochastic programming. Kybernetika *51* (2015), 433–456. DOI:10.14736/kyb-2015-3-0433

[10] S. T. Rachev and W. Römisch: Quantitative stability and stochastic programming: the method of probabilistic metrics. Math. Oper. Res. *27* (2002), 792–818. DOI:10.1287/moor.27.4.792.304

[11] S. Shafieezadeh-Abadeh and P. M. Esfahani: Regularization via mass transportation. J. Machine Learning Res. *20* (2019), 1–68.

[12] A. Shapiro and H. Xu: Stochastic mathematical programs with equilibrium constrains, modeling and sample average approximation. Optimization *57* (2008), 395–418. DOI:https://doi.org/10.1080/02331930801954177

[13] A. N. Tikhonov and V. Y. Arsenin: Solutions of Ill-posed Problems. Winston and Sons, Washington DC 1977.

[14] V. N. Vapnik: Statistical Learning Theory. Wiley and Sons, New York 1998.

[15] V. Vapnik and R. Izmailov: Synergy of monotonic rules. J. Machine Learning Res. *17* (2016), 988–999.

Evgueni Gordienko, Departamento de matematicas, UAM-I. Avenida San Rafael Atlixco 186, Col. Vicentina 09340. Mexico City. Mexico.
e-mail: gord@ xanum.uam.mx

Yury Gryazin, Department of Mathematics and Statistics, Idaho State University, 921 South 8th Ave, Stop 8085, Pocatello, ID 83209. U. S. A.
e-mail: gryazin@isu.edu