

**A NOTE ON THE GRADIENT OF
THE INTEGRAL ENTROPY FUNCTION**

Mengwei Xu¹ §, Liwei Zhang²

¹School of Mathematical Sciences
Dalian University of Technology
Dalian, 116024, P.R. CHINA

²Dalian University of Technology
Dalian, 116024, P.R. CHINA

Abstract: In this paper, we consider a parametric minimization problem and use the integral entropy function to approximate its value function. We show that the repeated limits of the gradient of the integral entropy function may be strictly contained in the Clarke generalized gradient of the value function in some cases. We also apply the result to bilevel programs.

AMS Subject Classification: 65K10, 90C26

Key Words: integral entropy function, value function, bilevel program

1. Introduction

The parametric minimization is a problem in m variables that depends on n parameters which can be specified by

$$\min_{y \in Y} f(x, y)$$

for each vector $x \in X$, where $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$ are closed convex sets and $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is continuously differentiable.

Received: May 19, 2014

© 2014 Academic Publications, Ltd.
url: www.acadpubl.eu

§Correspondence author

We denote the value function of the parametric minimization problem by

$$V(x) := \inf_{y \in Y} f(x, y).$$

While the major difficulty encountered is the nondifferentiability of the value function. [5] proposed to use the integral entropy function

$$\begin{aligned} \gamma_\rho(x) &:= -\rho^{-1} \ln \left(\int_Y \exp[-\rho f(x, y)] dy \right) \\ &= V(x) - \rho^{-1} \ln \left(\int_Y \exp[-\rho(f(x, y) - V(x))] dy \right) \end{aligned}$$

to approximate the value function and also showed the uniformly convergence as $\rho \rightarrow \infty$. [6, Theorem 5.1 and 5.5] showed that $\{\gamma_\rho(x) : \rho > 0\}$ is a smoothing function of $V(x)$, i.e., for any $x \in X$,

$$\lim_{z \rightarrow x, \rho \uparrow \infty} \gamma_\rho(z) = V(x)$$

and satisfies the gradient consistency property, i.e., for any $x \in X$,

$$\emptyset \neq \limsup_{z \rightarrow x, \rho \uparrow \infty} \nabla \gamma_\rho(z) \subseteq \partial V(x).$$

In this paper, we discuss the gradient of the integral entropy function.

2. Preliminaries

In this section, we present some background materials which will be used later on.

We first adopt the following standard notation in this paper. Given a function $G : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we denote its Jacobian by $\nabla G(z) \in \mathbb{R}^{m \times n}$ and, if $m = 1$, the gradient $\nabla G(z) \in \mathbb{R}^n$ is considered as a column vector. For a set $\Omega \subseteq \mathbb{R}^n$, we denote by $\text{int } \Omega$, $\text{ri } \Omega$, $\text{co } \Omega$, and $\text{dist}(x, \Omega)$ the interior, relative interior, the convex hull, and the distance from x to Ω respectively. In addition, we let \mathbf{N} be the set of nonnegative integers and $\exp[z]$ be the exponential function.

Let $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ be Lipschitz continuous near \bar{x} . The Clarke generalized directional derivative of φ at \bar{x} in direction d is defined by

$$\varphi^\circ(\bar{x}; d) := \limsup_{x \rightarrow \bar{x}, t \downarrow 0} \frac{\varphi(x + td) - \varphi(x)}{t}.$$

The Clarke generalized gradient of φ at \bar{x} is a convex and compact subset of \mathbb{R}^n defined by

$$\partial\varphi(\bar{x}) := \{\xi \in \mathbb{R}^n : \xi^T d \leq \varphi^\circ(\bar{x}; d), \quad \forall d \in \mathbb{R}^n\}.$$

Detailed discussions of the Clarke generalized gradient and its properties can be found in [2, 3].

For a nonempty closed set $\Omega \subseteq \mathbb{R}^n$ and a point $\bar{x} \in \Omega$, the Clarke tangent cone [2, 3] of Ω at \bar{x} is given by

$$\mathcal{T}_\Omega(\bar{x}) := \{d \in \mathbb{R}^n : \text{dist}^o(x^*, \Omega) = 0\},$$

and the Clarke normal cone [2, 3] of Ω at \bar{x} is given by

$$\mathcal{N}_\Omega(\bar{x}) := \{\zeta \in \mathbb{R}^n : \zeta^T d \leq 0, \quad \forall d \in \mathcal{T}_\Omega(\bar{x})\}$$

respectively.

Proposition 2.1 (Danskin’s Theorem). ([3, Page 99] or [4]) *Let $Y \subseteq \mathbb{R}^m$ be a compact set and $f(x, y)$ be a function defined on $\mathbb{R}^n \times \mathbb{R}^m$ that is continuously differentiable at \bar{x} . Then the value function*

$$V(x) := \min\{f(x, y) : y \in Y\}$$

is Lipschitz continuous near \bar{x} and its Clarke generalized gradient at \bar{x} is

$$\partial V(\bar{x}) = \text{co}\{\nabla_x f(\bar{x}, y) : y \in S(\bar{x})\}, \tag{2.1}$$

where $S(\bar{x})$ is the set of all minimizers of $f(x^*, y)$ over $y \in Y$.

In the end of this section, we review the definition of Lebesgue Measure[10].

Definition 2.1 (Exterior measure). If $E \subseteq \mathbb{R}^n$, the exterior measure of E is

$$m_*(E) := \inf \sum_{j=1}^{\infty} |Q_j|,$$

where $|Q|$ denotes the volume of a closed cube Q and the infimum is taken over all countable closed cubes $\{Q_j\}_{j=1}^{\infty}$ such that $\cup_{j=1}^{\infty} Q_j \supseteq E$.

Definition 2.2 (Lebesgue measurability). A set $E \subseteq \mathbb{R}^n$ is Lebesgue measurable if, for any $\epsilon > 0$, there exists an open set O with $E \subseteq O$ and

$$m_*(O - E) \leq \epsilon.$$

For a measurable set E , $m^*(E)$ is called the Lebesgue measure of E .

3. The Gradient of the Integral Entropy Function

The following lemma is a directly result of Theorem 1.17 (b) [9].

Lemma 3.1. *For any point $x^* \in X$, we have*

$$\limsup_{x \rightarrow x^*} S(x) \subseteq S(x^*). \tag{3.1}$$

Lemma 3.2. [6, Theorems 5.2, 5.3] *For any $\rho > 0$,*

$$\nabla \gamma_\rho(x) = \int_Y \mu_\rho(x, y) \nabla_x f(x, y) dy,$$

where

$$\mu_\rho(x, y) := \frac{\exp[-\rho f(x, y)]}{\int_Y \exp[-\rho f(x, y)] dy}$$

and for any $x \in X$,

$$\lim_{\rho \rightarrow \infty} \mu_\rho(x, y) = \begin{cases} m^*(S(x))^{-1}, & y \in S(x), \\ 0, & y \in Y \setminus S(x). \end{cases}$$

Here, $m^*(S(x))^{-1} := +\infty$ if $m^*(S(x)) = 0$.

Theorem 3.1. *Let x^* be a limiting point of $\{x_k\}$ and $\rho_k \rightarrow \infty$, if $S(x^*) = \{y^*\}$, then we have*

$$\lim_{k \rightarrow \infty} \nabla \gamma_{\rho_k}(x_k) = \nabla_x f(x^*, y^*).$$

Proof. From the gradient consistent property of $\gamma_\rho(\cdot)$, for any vector v and subset $K \subset \mathbf{N}$ such that

$$v := \lim_{k \rightarrow \infty, k \in K} \nabla \gamma_{\rho_k}(x_k),$$

we have $v \in \partial V(x^*)$. By Danskin’s Theorem, $\partial V(x^*) = \{\nabla_x f(x^*, y^*)\}$ since $S(x^*) = \{y^*\}$. Since v and K are arbitrary accumulation point and subset, $\partial V(x^*)$ is a singleton set, we have $\lim_{k \rightarrow \infty} \nabla \gamma_{\rho_k}(x_k) = \nabla_x f(x^*, y^*)$. □

Theorem 3.2. *For any point x^* and any subset $X_0 \subseteq X$ such that $x^* \in cl X_0$ and $m^*(S(x)) \neq 0, \forall x \in X_0$. We have*

$$\limsup_{x \rightarrow x^*, x \in X_0} \lim_{\rho \rightarrow \infty} \nabla \gamma_\rho(x) \subseteq U(x^*) := co\{\nabla_x f(x^*, y) : y \in S_0^*\}, \tag{3.2}$$

where $S_0^* = \limsup_{x \rightarrow x^*, x \in X_0} S(x)$.

Proof. For any $x \in X_0$ and $y \in S(x)$, we have $f(x, y) - V(x) = 0$. Thus, by the definition of $\mu_\rho(x, y)$,

$$\begin{aligned} \mu_\rho(x, y) &= \frac{\exp[-\rho(f(x, y) - V(x))]}{\int_Y \exp[-\rho(f(x, z) - V(x))] dz} \\ &= \frac{\exp[-\rho(f(x, y) - V(x))]}{\int_{S(x)} \exp[-\rho(f(x, z) - V(x))] dz + \int_{Y \setminus S(x)} \exp[-\rho(f(x, z) - V(x))] dz} \\ &= \frac{1}{m^*(S(x)) + \int_{Y \setminus S(x)} \exp[-\rho(f(x, z) - V(x))] dz} \\ &\leq \frac{1}{m^*(S(x))}. \end{aligned}$$

Let

$$M(x, y) = \begin{cases} m^*(S(x))^{-1} |\nabla_x f(x, y)|, & y \in S(x), \\ c |\nabla_x f(x, y)|, & y \in Y \setminus S(x). \end{cases}$$

$\int_{y \in Y} M(x, y) dy < \infty$ from the continuity of $\nabla_x f(x, \cdot)$. From Lemma 3.2, there exist $\bar{\rho}$ and a constant $c > 0$ such that for $\rho > \bar{\rho}$ and $y \in Y \setminus S(x)$, $\mu_\rho(x, y) \leq c$. Thus for any $x \in X_0$ and $\rho > \bar{\rho}$, $|\mu_\rho(x, y) \nabla_x f(x, y)| \leq M(x, y)$.

For any $v \in \limsup_{x \rightarrow x^*, x \in X_0} \lim_{\rho \rightarrow \infty} \nabla \gamma_\rho(x)$, we assume there exists a subset $X_1 \subseteq X_0$ such that $v = \lim_{x \rightarrow x^*, x \in X_1} \lim_{\rho \rightarrow \infty} \nabla \gamma_\rho(x)$. From the Lebesgue convergence theorem, we have

$$\begin{aligned} &\lim_{x \rightarrow x^*, x \in X_1} \lim_{\rho \rightarrow \infty} \nabla \gamma_\rho(x) \\ &= \lim_{x \rightarrow x^*, x \in X_1} \int_Y \lim_{\rho \rightarrow \infty} \mu_\rho(x, y) \nabla_x f(x, y) dy \\ &= \lim_{x \rightarrow x^*, x \in X_1} \int_{S(x)} \frac{1}{m^*(S(x))} \nabla_x f(x, y) dy. \end{aligned} \tag{3.3}$$

Since both X and Y are compact and $\nabla_x f(x, y)$ is continuous on $X \times Y$, $\nabla_x f(x, y)$ is uniformly continuous on $X \times Y$. Thus, for any $\epsilon > 0$, there exists $\delta > 0$ such that, for any (z_1, y_1) and (z_2, y_2) satisfying $\|(z_1, y_1) - (z_2, y_2)\| \leq 3\delta$,

$$\|\nabla_x f(z_1, y_1) - \nabla_x f(z_2, y_2)\| \leq \epsilon. \tag{3.4}$$

Furthermore, there exists a constant $c_1 > 0$ such that $|\nabla_x f(x^*, y)| \leq c_1$ for all $y \in S(x^*)$.

Due to the fact that S_0^* is compact, there exists $0 < \hat{\delta} < \frac{\delta}{2}$ such that $\bigcup_{y \in S_0^*} B(y, \frac{\delta}{2}) \supseteq (S_0^* + \hat{\delta}\mathbf{B})$, where \mathbf{B} denotes the closed unit ball. We get from the Heine-Borel covering theorem that there exist $N > 0$ and $y_i^* \in S_0^*$ such that

$$\bigcup_{i=1}^N B(y_i^*, \frac{\delta}{2}) \supseteq (S_0^* + \hat{\delta}\mathbf{B}).$$

From the Lemma 3.1 and the Theorem 4.10 (b) [9], we know that for $\hat{\delta} > 0$, $\forall x \in X_0$ such that $\|x - x^*\|$ sufficiently small, we have

$$S(x) \subseteq S_0^* + \hat{\delta}\mathbf{B}.$$

Thus there exists an index set $I(x) \subseteq \{1, \dots, N\}$ such that for each $i \in I(x)$, $B(y_i^*, \frac{\delta}{2}) \cap S(x) \neq \emptyset$ from Lemma 3.1. For each $i \in I(x)$, there exists $y_i^x \in B(y_i^*, \frac{\delta}{2}) \cap S(x)$. Note that $I(x)$ may not equal to the whole index set $\{1, \dots, N\}$ since S_0^* is the outer limit.

Let $\hat{B}(y_i^x, \delta) = B(y_i^x, \delta) \cap S(x)$, $i \in I(x)$, for each $x \in X_0$ such that $\|x - x^*\|$ sufficiently small. Without loss of generality, we assume that $I(x) = \{1, \dots, l_x\}$, $l_x \leq N$. Let $\Omega_1 := \hat{B}(y_1^x, \delta)$, $\Omega_i := \hat{B}(y_i^x, \delta) \setminus \hat{B}(y_i^x, \delta) \cap (\bigcup_{j=1}^{i-1} \hat{B}(y_j^x, \delta))$ for $i = 2, \dots, l_x$. For all $i \in I(x)$,

$$B(y_i^x, \delta) \supseteq B(y_i^*, \frac{\delta}{2}).$$

It is obvious that $\Omega_1 \cap \dots \cap \Omega_{l_x} = \emptyset$ and $\bigcup_{i=1}^{l_x} \Omega_i = S(x)$.

For any $x \in X_1$, let $\lambda_i^x := \frac{m^*(\Omega_i)}{m^*(S(x))}$, $i = 1, \dots, l_x$. It follows that $\lambda_i^x \geq 0$ and $\sum_{i=1}^{l_x} \lambda_i^x = 1$. Note that $\forall y \in \Omega_i$, $\|y - y_i^*\| \leq 1.5\delta$. Thus for $x \in X_1$ such that $\|x - x^*\| < \delta$, we have

$$\begin{aligned} & \left\| \int_{S(x)} \frac{1}{m^*(S(x))} \nabla_x f(x, y) dy - \sum_{i=1}^{l_x} \lambda_i^x \nabla_x f(x^*, y_i^*) \right\| \\ & \leq \left\| \sum_{i=1}^{l_x} \int_{\Omega_i} \frac{1}{m^*(S(x))} (\nabla_x f(x, y) - \nabla_x f(x^*, y_i^*)) dy \right\| \\ & \leq \epsilon \sum_{i=1}^{l_x} \frac{m^*(\Omega_i)}{m^*(S(x))} \\ & = \epsilon. \end{aligned}$$

Thus

$$\text{dist}(\lim_{\rho \rightarrow \infty} \nabla \gamma_\rho(x), U(x^*)) \rightarrow 0, \text{ as } x \rightarrow x^*, x \in X_1.$$

We complete the proof. □

Obviously, for any x^* , we have $U(x^*) \subseteq \partial V(x^*)$.

4. Applications to Bilevel Program

Many scientific problems such as a very important model in economics called the principal-agent problem [7] can be formulated as the following simple bilevel program:

$$\begin{aligned} \text{(SBP)} \quad & \min && F(x, y) \\ & \text{s.t.} && y \in S(x), \\ & && x \in X, \end{aligned}$$

where $S(x)$ denotes the set of solutions of the lower level program

$$\text{(P}_x\text{)} \quad \min_{y \in Y} f(x, y),$$

where $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$ are closed convex sets and $F, f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ are twice continuously differentiable.

For a numerical purpose, Outrata [8] proposed to reformulate (SBP) as the following single level optimization problem:

$$\begin{aligned} \text{(VP)} \quad & \min && F(x, y) \\ & \text{s.t.} && f(x, y) - V(x) \leq 0, \\ & && (x, y) \in X \times Y. \end{aligned} \tag{4.1}$$

Since the value function $V(x)$ is generally nonsmooth even when the function $f(x, y)$ is smooth, the problem (VP) is a nonsmooth problem. To cope with such difficulty, [6] approximated the value function by its integral entropy function and proposed a smoothing gradient projection algorithm to solve (VP).

From the definition of $V(x)$, for any $(x, y) \in X \times Y$, we always have $f(x, y) - V(x) \geq 0$. Hence for any feasible point (x^*, y^*) of the problem (VP), (x^*, y^*) is an optimal solution of the problem

$$\min_{(x,y) \in X \times Y} f(x, y) - V(x),$$

which deduces that

$$0 \in \nabla f(x^*, y^*) - \partial V(x^*) \times \{0\} + \mathcal{N}_{X \times Y}(x^*, y^*). \quad (4.2)$$

Therefore, the GMFCQ never holds for the problem (VP) and hence the non-smooth KKT condition may not hold at a local optimal solution.

While for a sequence of iteration points $\{(x_k, y_k)\}$ which convergent to (x^*, y^*) , the set $\limsup_{k \rightarrow \infty} \nabla \gamma_{\rho_k}(x^k)$ may strictly contain in $\partial V(x)$. Therefore while (4.2) holds, the following inclusion may not hold:

$$0 \in \nabla f(x^*, y^*) - \limsup_{k \rightarrow \infty} \nabla \gamma_{\rho_k}(x^k) \times \{0\} + \mathcal{N}_{X \times Y}(x^*, y^*), \quad (4.3)$$

which guarantees a sequence of bounded multipliers and thus (x^*, y^*) is a stationary point of the problem (VP) [11].

In the case where $S(x^*)$ is a singleton set, we know that the condition (4.3) always holds from Theorem 3.1. While in this case, the value function is continuously differentiable around x^* and we suggest to use the first order approach which replace the lower level program by its Kurash-Kuhn-Tucker (KKT) condition to solve such problem.

For any feasible point (x^*, y^*) , since S_0^* is possibly a proper subset of $S(x^*)$, which implies that $U(x^*)$ is strictly contained in $\partial V(x^*)$, the following inclusion may fails:

$$0 \in \nabla f(x^*, y^*) - U(x^*) \times \{0\} + \mathcal{N}_{X \times Y}(x^*, y^*)$$

although the condition (4.2) always holds. If one can show the relationship between $\limsup_{x \rightarrow x^*, x \in X_0} \lim_{\rho \rightarrow \infty} \nabla \gamma_{\rho}(x)$ and $\limsup_{\rho \rightarrow \infty, x \rightarrow x^*, x \in X_0} \nabla \gamma_{\rho}(x)$, then the condition (4.3) can be discussed accordingly. This remains a topic of our future research.

Acknowledgments

The research of this author was supported by the National Natural Science Foundation of China under projects No. 11071029, No. 91330206 and No. 91130007.

References

- [1] X. Chen, R.S. Womersley and J.J. Ye, *Minimizing the condition number of a gram matrix*, SIAM J. Optim., **21**(2011), 127–148.

- [2] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
- [3] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern and P.R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer, New York, 1998.
- [4] J.M. Danskin, *The Theory of Max-Min and its Applications to Weapons Allocation Problems*, Springer, New York, 1967.
- [5] S.C. Fang and S.Y. Wu, *Solving min-max problems and linear semi-infinite programs*, *Comput. Math. Appl.*, **32**(1996), 87-93.
- [6] G.H. Lin, M. Xu and J.J. Ye, *On solving simple bilevel programs with a nonconvex lower level program*, *Math. Program., series A*, **144**(2014), 277-305.
- [7] J. Mirrlees, *The theory of moral hazard and unobservable behaviour: Part I*, *Rev. Econ. Stud.*, **66**(1999), 3-22.
- [8] J.V. Outrata, *On the numerical solution of a class of Stackelberg problems*, *Z. Oper. Res.*, **34**(1990), 255-277.
- [9] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.
- [10] E.M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, Springer, 2005.
- [11] M. Xu, J.J. Ye and L. Zhang *Smoothing SQP methods for solving degenerated nonsmooth constrained optimization problems with applications to bilevel programs*, ArXiv:1403.1636 [math.OC], 2014.

