On max-k-sums

Michael J. Todd

January 10, 2018

School of Operations Research and Information Engineering, Cornell University

http://people.orie.cornell.edu/∼miketodd/todd.html

11th US-Mexico Workshop on Optimization and its Applications, Huatulco, January 2018

1. Definitions

Given scalars $y_1, \ldots, y_n \in I\!R$, define their max-$k$-sum as

$$M^k(y) := \max_{|K|=k} \sum_{i \in K} y_i = \sum_{j=1}^{k} y_{[j]}$$

and their min-k-sum as

$$m^k(y) := \min_{|K|=k} \sum_{i \in K} y_i = \sum_{j=n-k+1}^{n} y_{[j]},$$

where $y_{[1]}, \ldots, y_{[n]}$ denote the y_i's in nonincreasing order.

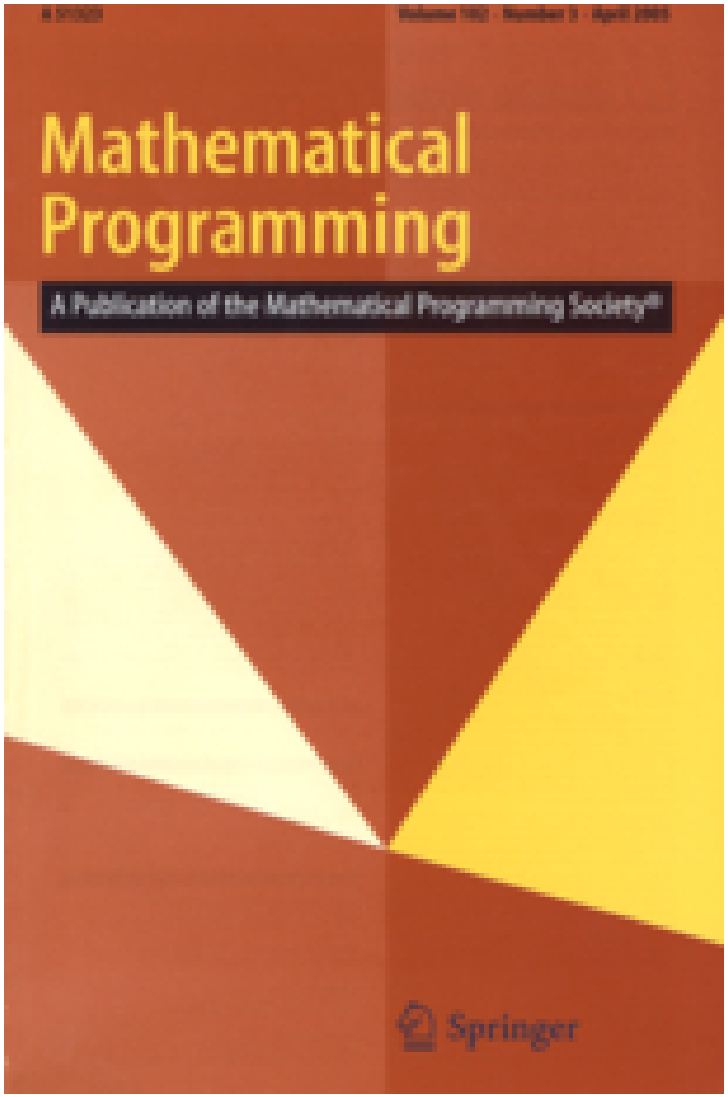
These arise in

- constraints in scenario-based conditional value at risk computation (giving a convex problem; restricting k out of n gives a MIP),
- penalties for peak demand in electricity modelling,
- and are related to Owl norms used in regularization in machine learning problems.

Given functions $f_1, \ldots, f_n$ on $I\!R^d$, define

$$F^k(t) := M^k(f_1(t), \ldots, f_n(t)) \text{ and}$$

$$f^k(t) := m^k(f_1(t), \ldots, f_n(t)).$$


Mathematical
Programming
A Publication of the Mathematical Programming Society®
Springer

2. Two Questions

a) How can we define

- smooth approximations to F^k and f^k,

maintaining certain properties of the unsmoothed functions?

b) How can we define (original or smoothed) max-k-sums [min-k-sums] if

- the y_i's lie in a vector space ordered by a convex cone,

again preserving properties of the real case?

Note that F^k (f^k) is the composition of M^k (m^k) with the map f from t to $(f_1(t), \ldots, f_n(t))$, so most of the time we address only the latter functions.

Desirable Properties

- 0-consistency: $M^0(y) = m^0(y) = 0$;
- n-consistency: $M^n(y) = m^n(y) = \sum_i y_i$;
- sign-reversal: $m^k(y) = -M^k(-y)$;
- summability: $M^k(y) + m^{n-k}(y) = \sum_i y_i$;
- translation invariance: $M^k(y + \eta\mathbf{1}) = M^k(y) + k\eta$, $m^k(y + \eta\mathbf{1}) = m^k(y) + k\eta$.
- scale invariance: for $\alpha > 0$, $M^k(\alpha y) = \alpha M^k(y)$, $m^k(\alpha y) = \alpha m^k(y)$.
- convexity: if $f_1, \ldots, f_n$ are convex, so is F^k; if they are concave, so is f^k.

3. Smoothing via Randomization in the Domain

A classical technique is to approximate a nonsmooth function h via a convolution or as an expectation:

$$\begin{aligned} \tilde{h}(t) &:= E_s h(t-s) \\ &= \int h(t-s)\phi(s)ds, \end{aligned}$$

where ϕ is the probability density function of a localized random variable $s \in \Re^d$.

However, this shrinks the domain dom $h := \{t : h(t) < \infty\}$, inappropriate in some cases, and requires a computationally burdensome d-dimensional integration.

4. A Modification

Instead, we randomize in the range of the functions: Let $\xi_1, \ldots, \xi_n$ be iid random variables distributed like the (continuous) random variable Ξ and set

$$\begin{aligned} \bar{M}^k(y) &:= E_{\xi_1,\ldots,\xi_n} \max_{|K|=k} \sum_{i\in K}(y_i - \xi_i) + kE\Xi \\ \bar{m}^k(y) &:= E_{\xi_1,\ldots,\xi_n} \min_{|K|=k} \sum_{i\in K}(y_i - \xi_i) + kE\Xi \end{aligned}$$

and then $\bar{F}^k(t) := \bar{M}^k(f(t))$ and $\bar{f}^k(t) := \bar{m}^k(f(t))$. These functions inherit the smoothness of the f_i's. Moreover, they inherit the domains of the nonsmooth functions. Further, they satisfy 0- and n-consistency, summability, translation invariance, and convexity, and the approximation bounds

$$M^k(y) \le \bar{M}^k(y) \le M^k(y) + \bar{M}^k(0) \le M^k(y) + \min(k\bar{M}^1(0), -(n-k)\bar{m}^1(0))$$

and

$$m^k(y) \ge \bar{m}^k(y) \ge m^k(y) + \bar{m}^k(0) \ge m^k(y) - \min((n-k)\bar{M}^1(0), -k\bar{m}^1(0)).$$

They do not satisfy sign reversal or scale invariance, but

$$\bar{m}^k(y;\Xi) = -\bar{M}^k((-y;-\Xi)$$

and

$$\bar{M}^k(\alpha y;\alpha\Xi) = \alpha\bar{M}^k(y;\Xi),$$

and similarly for $\bar{m}^k$, for positive α.

5. Evaluation

To enable fairly efficient evaluation, we choose Gumbel random variables:
$P(\Xi > x) = \exp(-\exp(x))$, $E\Xi = -\gamma$.
Recall that $z_{[k]}$ denotes the kth largest component of a vector $z \in \Re^n$. We are interested in
$q_k := E((y - \xi)_{[k]})$.

$$q_k = \cdots = \sum_{|K|<k} (-1)^{k-|K|-1} \binom{n- \mid K \mid -1}{k- \mid K \mid -1} \ln \sum_{h \notin K} \exp(y_h) + \gamma.$$

From this, we obtain

Theorem 1

$$\bar{M}^k(y) = \sum_{|K|<k} (-1)^{k-|K|-1} \binom{n- \mid K \mid -2}{k- \mid K \mid -1} \ln \sum_{h \notin K} \exp(y_h).$$

□

(Here $\binom{0}{0} := \binom{-1}{0} := 1$, and otherwise $\binom{p}{q} := 0$ if $p < q$.)

We have reduced the work from an n-dimensional integration to a sum over $O((n)^{k-1})$ terms.

Note that almost all the terms disappear for $k = n$, and we get $\bar{M}^n(y) = M^n(y)$ as expected.

6. Examples

$k = 1$: Here only $K = \emptyset$ contributes to the sum, so we obtain

$$\bar{M}^1(y) = \ln\left(\sum_h \exp(y_h)\right).$$

Such functions have been used as potential functions in theoretical computer science, starting with Shahrokhi-Matula and Grigoriadis-Khachiyan, and are discussed by Tunçel and Nemirovski in the context of barrier functions.

They also appear in the economic literature on consumer choice, dating back to the 1960s (e.g., Luce and Suppes).

This function is sometimes called the soft maximum of the y_j's. This term is also used for the weight vector

$$\left(\frac{\exp(y_i)}{\sum_h \exp(y_h)}\right).$$

Note that this is the gradient of $\bar{M}^1$ and thus the gradient of $\bar{F}^1$ is the weighted combination of those of the f_j's using these weights for $y = f(t)$.

$k = 2$: Here K can be the empty set or any singleton, and we find

$$\begin{aligned}\bar{M}^2(y) &= -(n-2)\ln\left(\sum_h \exp(y_h)\right) + \sum_i \ln\left(\sum_{h\neq i} \exp(y_h)\right) \\ &= \ln\left(\sum_{h\neq 2} \exp(y_{[h]})\right) + \ln\left(\sum_{h\neq 1} \exp(y_{[h]})\right) + \\ &\quad \sum_{i>2} \ln\left(1 - \frac{\exp(y_{[i]})}{\sum_h \exp(y_h)}\right).\end{aligned}$$

Bounds

Theorem 2

$$M^k(y) \leq \bar{M}^k(y) \leq M^k(y) + k\ln n.$$

If we want a closer (but "rougher") approximation, we can scale the Gumbel random variables by $\alpha < 1$, or equivalently, scale the vector y by α^{-1}, apply the formulae above, and then scale the result by α.

If the y_i's differ by orders of magnitude, the above expressions need to be carefully evaluated, but at the same time, we may be able to ignore many of the terms.

7. Formulation via (Continuous) Optimization Problems

We note that $M^1(y)$ can be obtained as the optimal value of

$$P(M^1): \quad \min\{x : x \geq y_i \text{ for all } i\}$$

and

$$D(M^1): \quad \max\{\sum_i u_i y_i : \sum_i u_i = 1,\ u_i \geq 0 \text{ for all } i\};$$

either the smallest upper bound on the y_i's or their largest convex combination.

These are probably the simplest and most intuitive dual linear programming problems of all!

Analogously, $M^k(y)$ is the optimal value of

$$D(M^k): \quad \max\{\sum_i u_i y_i : \sum_i u_i = k,\ 0 \leq u_i \leq 1 \text{ for all } i\},$$

with feasible region $U := U^k$, whose dual is

$$P(M^k): \quad \min\{kx + \sum_i z_i : x + z_i \geq y_i, z_i \geq 0, \text{ for all } i\}.$$

(Note that there is a slight abuse of notation: for $k = 1$, these are not the same problems as above, but can be seen to be equivalent.)

We can similarly obtain $m^1(y)$ and $m^k(y)$.

8. Smoothing via Perturbation (à la Nesterov)

We define $\hat{M}^k(y)$ to be the optimal value of

$$\hat{D}(M^k): \quad \max\{\sum_i u_i y_i - g^*(u) : u \in U\},$$

where $g^* := g^{*k}$ is a strongly convex function on $U := U^k$ satisfying certain properties, $+\infty$ off $\{u : \sum_i u_i = k\}$, with minimum 0 and maximum Δ on U. We define $\hat{m}^k(y)$, $\hat{F}^k(t)$, and $\hat{f}^k(t)$ analogously.

We then have 0- and n-consistency, sign reversal, translation invariance, and summability as long as $g^{*n-k}(u) = g^{*k}(\mathbf{1} - u)$ for $u \in U^{n-k}$. Moreover, $\hat{M}^k$ is Lipschitz continuously differentiable. We also have scale invariance in the form $\hat{M}^k(\alpha y, \alpha g^*) = \alpha M^k(y, g^*)$, the convexity property for $\hat{F}^k$ and $\hat{f}^k$, and the bounds

$$M^k(y) - \Delta \le \hat{M}^k(y) \le M^k(y), \quad m^k(y) \le \hat{m}^k(y) \le m^k(y) + \Delta.$$

The dual of $\hat{D}(M^k)$ is

$$\hat{P}(M^k): \quad \min\{kx + \sum_i z_i + g(w) : x + z_i \ge y_i - w_i, z_i \ge 0, \text{ for all } i \quad (\text{and } \sum_i w_i = 0)\},$$

where g is the convex conjugate of g^*.

9. Examples

Quadratic function

Let

$$g^*(u) := g^{*k}(u) := \frac{\beta}{2}(\|u\|_2)^2 - \frac{\beta(k)^2}{2n}.$$

Then we can show that $\hat{D}(M^k)$ is solved by

$$u_i = \text{mid}(0, y_i/\beta - \lambda, 1) \text{ for all } i,$$

for some λ, and we can solve the problem in $O(n \ln n)$ time by sorting and a binary search.

Single-sided entropic function

Next we let

$$g^*(u) := g^{*k}(u) := \sum_i u_i \ln u_i + k \ln\left(\frac{n}{k}\right)$$

for nonnegative u_i's summing to k. Now we can find the optimal u from

$$u_i = \min(\exp(y_i - \lambda), 1) \text{ for all } i,$$

for some λ, so the problem can again be solved in $O(n \ln n)$ time by sorting and a binary search. Interestingly, $\hat{M}^1(y) = \bar{M}^1(y) - \ln n$, but there is no such relation for $k > 1$, and the $\hat{M}^k$'s are much easier to evaluate than the $\bar{M}^k$'s.

10. Max-k-Sums in General Spaces

Now suppose $y_1, \ldots, y_n$ lie in a finite-dimensional real vector space E ordered by a closed convex pointed cone $\mathcal{K}$ with nonempty interior. Let E^* denote the dual space, with dual cone $\mathcal{K}^* := \{u \in E^* : \langle u, x\rangle \geq 0 \text{ for all } x \in \mathcal{K}\}$. Then

$$x \succeq z, x, z \in E \text{ means } x - z \in \mathcal{K}, \quad u \succeq^* v, u, v \in E^* \text{ means } u - v \in \mathcal{K}^*.$$

We also write $z \preceq x$ and $v \preceq^* u$ with the obvious definitions.

We would like to define the max-k-sum and the min-k-sum of the y_i's in E, and smooth approximations to them, to conform with their definitions in $I\!R$. We write $((y_i))$ for $(y_1, \ldots, y_n) \in E^n$ for ease of notation.

Our prime examples for E and $\mathcal{K}$ are:

- $I\!R$ and $I\!R_+$;
- $I\!R^p$ and $I\!R^p_+$;
- the space of real (complex) symmetric (Hermitian) $d \times d$ matrices, and the cone of positive semidefinite matrices; and
- $I\!R^{1+p}$ and the second-order cone $\{(\xi; x) \in I\!R^{1+p} : \xi \geq \|x\|_2\}$.

Some results below hold just for symmetric cones — all those above are symmetric.

11. "Smoothing via Randomization"

This makes no sense, since we don't yet know how to define the max-k-sum to add randomization to!

But we can use the formulae we derived for the case of reals if $\exp$ and $\ln$ are defined. And they are for symmetric cones!

For example, for symmetric matrices, if $A = VDV^T$ is the eigenvalue decomposition of A, then $\exp(A) = V \exp(D) V^T$, and if A is positive definite, $\ln(A) = V \ln(D) V^T$. Here $\exp$ and $\ln$ are defined for diagonal matrices by applying the scalar version to each diagonal entry.

We can show that

$$\ln\left(\sum \exp(y_i)\right) \succeq y_j$$

for each j (but not a similar result for $k = 2$).

These formulae satisfy translation invariance for $((y_i + \eta e))$, where e is the unit element in the symmetric cone.

12. Definition via Optimization Formulations

If we directly translate $P(M^k)$ to this setting, we find the objective function is not a scalar, so we choose $v \in \text{int}(\mathcal{K}^*)$ and then define

$$P(M^k((y_i)) : \quad \min\{k\langle v, x\rangle + \sum_i \langle v, z_i\rangle : x + z_i \succeq y_i,\ z_i \succeq 0, \text{ for all } i\}$$

and

$$D(M^k((y_i))) : \quad \max\{\sum_i \langle u_i, y_i\rangle : \sum_i u_i = kv,\ 0 \preceq^* u_i \preceq^* v \text{ for all } i\}.$$

with feasible region $U := U^k$ in E^{*n}.

We again choose a suitable strongly convex g^* on U, with convex conjugate g, and then define

$$\hat{P}(M^k((y_i))) : \quad \min\{k\langle v, x\rangle + \sum_i \langle v, z_i\rangle + g((w_i)) : x + z_i \succeq y_i - w_i, z_i \succeq 0, \text{ for all } i\}$$

and

$$\hat{D}(M^k((y_i))) : \quad \max\{\sum_i \langle u_i, y_i\rangle - g^*((u_i)) : \sum_i u_i = kv,\ 0 \preceq^* u_i \preceq^* v \text{ for all } i\}.$$

Our conditions on g^* imply that we can add the constraint $\sum_i w_i = 0$ without loss of generality.

Of course, the values of all these problems are scalars, and so will not provide the definitions we need. We therefore set

$$M^k((y_i)) := \{kx + \sum_i z_i : (x, (z_i)) \in \mathsf{Argmin}(P(M^k((y_i))))\}$$

and analogously $m^k((y_i))$ using Argmax.

(Here Argmin and Argmax denote the sets of all optimal solutions to the problem given.)

For the perturbed problems, we add the extra constraint $\sum_i w_i = 0$ to remove the ambiguity from x, and define

$$\hat{M}^k((y_i)) := \{kx + \sum_i z_i : (x, (z_i), (w_i)) \in \mathsf{Argmin}(\hat{P}(M^k((y_i)))), \sum_i w_i = 0\}$$

and analogously $\hat{m}^k((y_i))$.

13. Properties

These functions satisfy:

- 0- and n-consistency, in the sense that

$$M^0((y_i)) = \{0\}, \quad M^n((y_i)) = \{\sum_i y_i\},$$

  etc.;

- sign-reversal;

- summability, in the sense that

$$M^k((y_i)) = \{\sum_i y_i\} - m^{n-k}((y_i)),$$

  and if $g^{*n-k}((u_i)) := g^{*k}((v - u_i))$, similarly for $\hat{M}^k$ and $\hat{m}^{n-k}$;

- translation invariance for any $\eta \in E$;

- positive scaling invariance in the natural sense; and

- dominance: for any K of cardinality k and any $y \in M^k((y_i))$,

$$y \succeq \sum_{i \in K} y_i;$$

- respect of product structure: if $\mathcal{K}$ is a product of cones (and g is separable), then $M^k((y_i))$ (and $\hat{M}^k((y_i))$) are products of the M^k's (and $\hat{M}^k$'s) for the constituent cones.

Computation

To calculate $M^k((y_i))$ or $\hat{M}^k((y_i))$ requires the solution of a linear or convex conic programming problem.

One case is easier: if the cone is symmetric and $n = 2$, $k = 1$, we have

$$M^1(y_1, y_2) = \{(y_1 + y_2)/2 + \mathsf{abs}((y_1 - y_2)/2)\},$$

where abs is defined using the eigenvalue decomposition like exp and ln.

Remarks

- Simple arguments show that $M^k((y_i))$ may not be a singleton, and may depend on v;
- An alternative way to define $M^1((y_i))$ is as the limit (if it exists)

$$\lim_{\alpha \downarrow 0} \alpha \ln \left(\sum_i \exp(y_i/\alpha) \right).$$

This does not agree with the definition above.

14. Conclusions

The simple max-k-sum can be smoothed either by randomization or by perturbing an optimization formulation of the function.

The latter approach suggests a way to generalize the function to the case of general cones.

Final remark: Contrary to God, Kronecker, and Backus, k need not be an integer in the second approach, and the same properties hold!

All the best, Don!