Projection onto Minkowski Sums with Application to Constrained Learning: Supplement

Kenneth Lange [1] **Joong-Ho Won** [2] **Jason Xu** [3]

A. Additional Application of Projection onto Minkowski Sums

Constraint relaxation If B_r is a ball $\{\boldsymbol{x} : \|\boldsymbol{x}\| \le r\}$ with respect to a norm $\|\cdot\|$, the sum $C_i + B_r$ is a kind of halo around C_i, yielding a relaxation of the constraint C_i. Within any projection-based method for solving constrained problems, one may replace a given C_i by its Minkowski sum $C_i + B_r$ to account for a degree of error encoded by r. In Bayesian methods, priors sharply constrained to a support set C_i lead to computational issues. Recent work considers posterior projections (Patra & Dunson, 2018) and relaxing such constraints via "d-expansion" (Duan et al., 2018), a concept that coincides exactly with the Minkowski sum described above.

B. Additional Definitions

Definition B.1 (Strongly convex and smooth functions)**.** *Let $C \in \mathbb{R}^d$ is a convex set and $\|\cdot\|$ be a norm over the smallest vector space containing C. We say function $f : C \to \mathbb{R}$ is α-strongly convex with respect to $\|\cdot\|$ if*

$$f(\boldsymbol{y}) \ge f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2,$$

and is β-smooth with respect to $\|\cdot\|$ if

$$f(\boldsymbol{y}) \le f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\beta}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$$

for all $\boldsymbol{x}, \boldsymbol{y} \in C$.

Definition B.2 (Domain of an extended real-valued function)**.** *Let $\psi : \mathbb{R}^d \mapsto \mathbb{R} \cup \{\infty\}$ be an extended real-valued function. The domain of ψ is defined by $\mathbf{dom}(\psi) = \{\boldsymbol{x} \in \mathbb{R}^d : \psi(\boldsymbol{x}) < \infty\}$. If $\mathbf{dom}(\psi) \ne \emptyset$, then ψ is called proper.*

C. Projection for Internal and Boundary Points

Though we are naturally interested in external points, for completeness here we discuss the case when $\boldsymbol{x}$ is an internal or boundary point of the Minkowski sum. If $\boldsymbol{x} \in A + B$, then finding appropriate summands $\boldsymbol{a}$ and $\boldsymbol{b}$ also succumbs to block descent. As already noted in the main text, the decomposition $\boldsymbol{x} = \boldsymbol{a} + \boldsymbol{b}$ is not necessarily unique. In practical examples, convergence of block descent can be exceedingly slow for an internal or boundary point of $A + B$. For this reason, it is useful to explore alternatives. One possibility is to minimize the proximity function

$$f(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}\|_2^2 + \frac{\rho}{2}\,\mathrm{dist}(\boldsymbol{a}, A)^2 + \frac{\rho}{2}\,\mathrm{dist}(\boldsymbol{b}, B)^2,$$

whose minimal value is 0 for any $\boldsymbol{x} \in A + B$ and $\rho > 0$. The MM principle (Lange, 2016) summarized in the section below suggests minimizing the surrogate function

$$g(\boldsymbol{a}, \boldsymbol{b} \mid \boldsymbol{a}_n, \boldsymbol{b}_n) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{a} - P_A(\boldsymbol{a}_n)\|_2^2 + \frac{\rho}{2}\|\boldsymbol{b} - P_B(\boldsymbol{b}_n)\|_2^2$$

to generate improved values $\boldsymbol{a}_{n+1}$ and $\boldsymbol{b}_{n+1}$. The stationarity conditions for the surrogate read

$$\mathbf{0} \quad = \quad -(\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}) + \rho[\boldsymbol{a} - P_A(\boldsymbol{a}_n)]$$

$$\mathbf{0} \quad = \quad -(\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}) + \rho[\boldsymbol{b} - P_B(\boldsymbol{b}_n)].$$

One can readily verify that this linear system has the solution

$$\begin{aligned}
\boldsymbol{a}_{n+1} &= \frac{1}{2+\rho}[\boldsymbol{x} - P_B(\boldsymbol{b}_n)] + \frac{1+\rho}{2+\rho}P_A(\boldsymbol{a}_n) \\
\boldsymbol{b}_{n+1} &= \frac{1}{2+\rho}[\boldsymbol{x} - P_A(\boldsymbol{a}_n)] + \frac{1+\rho}{2+\rho}P_B(\boldsymbol{b}_n).
\end{aligned}$$

These updates are guaranteed to reduce the objective $f(\boldsymbol{a}, \boldsymbol{b})$. It is straightforward to prove that the update map is nonexpansive when A and B are both convex. The objective $f(\boldsymbol{a}, \boldsymbol{b})$ is also convex in this setting, and stationary points and global minima coincide. Furthermore, any fixed point $(\boldsymbol{a}, \boldsymbol{b})$ with $\boldsymbol{a} \in A$ and $\boldsymbol{b} \in B$ satisfies $\boldsymbol{x} = \boldsymbol{a} + \boldsymbol{b}$. This algorithm is a special case of a class of algorithms called proximal distance algorithms (Xu et al., 2017), with the distinction that the tuning constant ρ need not be sent to ∞ when $\boldsymbol{x} \in A + B$.

D. Majorization-Minimization

The majorization-minimization (MM) principle provides a generic recipe for converting an optimization problem that is not immediately solvable (for instance, it may be non-convex or non-smooth) into a sequence of manageable problems. MM algorithms have become increasingly popular for large-scale optimization in statistics and machine learning (Lange, 2016), and includes expectation-maximization (EM) algorithms as a special case. MM algorithms operate by successively minimizing a sequence of *surrogate functions* $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$ majorizing the objective function $f(\boldsymbol{x})$ at the current iterate $\boldsymbol{x}_m$. The notion of majorization requires two conditions: tangency $g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) = f(\boldsymbol{x}_m)$ at the current iterate, and domination $g(\boldsymbol{x} \mid \boldsymbol{x}_m) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x}$. The update rule

$$\boldsymbol{x}_{m+1} := \arg\min_{\boldsymbol{x}} \; g(\boldsymbol{x} \mid \boldsymbol{x}_m)$$

implies the descent property

$$f(\boldsymbol{x}_{m+1}) \;\leq\; g(\boldsymbol{x}_{m+1} \mid \boldsymbol{x}_m) \;\leq\; g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) \;=\; f(\boldsymbol{x}_m).$$

Note that minimizing g is not strictly necessary: the weaker condition $g(\boldsymbol{x}_{m+1} \mid \boldsymbol{x}_m) \leq g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)$ also decreases $f(\boldsymbol{x})$. Maximizing a function can be accomplished by an analogous combination of sequential minorization and maximization.

E. Additional Simulation Results for the $\ell_{1,p}$-Overlapping Group Lasso

The full runtime comparison results of the simulation study of Sect. 5.1 of the main text is presented. Figure E.1 shows the runtime for varying dimension for fixed number of groups. The top right panel corresponds to the top left panel of Figure 1 in the main text. Figure E.1 illustrates the same information in a different format: runtime for varying number of groups for a fixed dimension. The bottom right panel corresponds to the top right panel of Figure 1 in the main text.

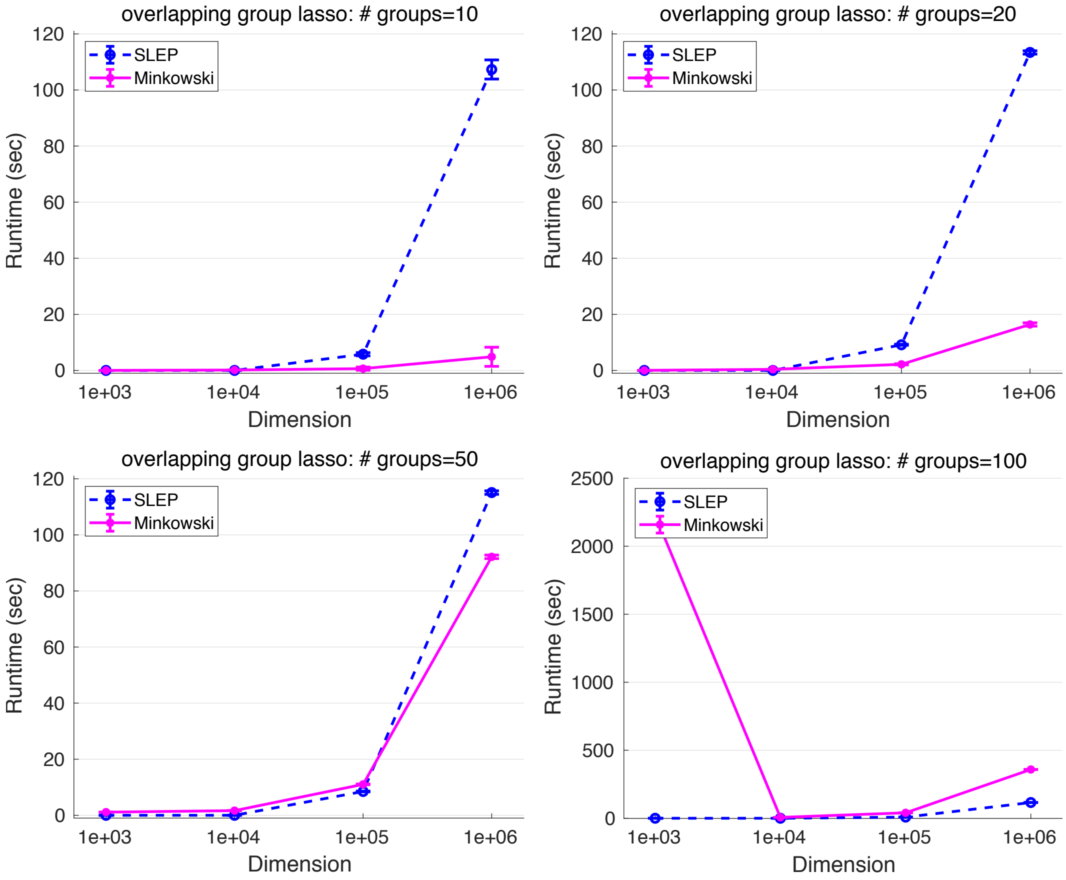

Figure E.1. Runtime-by-dimension comparison of the proposed Minkowski method and the dual projected gradient method by Yuan et al. (2011) for overlapping group lasso.

F. Additional Simulation Results for the Constrained Lasso

F.1. Additional Timing Results

The average runtime of path-following algorithm, Gurobi, ADMM, and the Minkowsi projection-based proximal gradient descent methods are shown in Figure F.1 for $\lambda/\lambda_{\max} = 0.4$ and 0.8. The results were obtained from the simulation as described in Sect. 5.2 of the main text, but omitted due to the space limit.

F.2. Accuracy

Following Gaines et al. (2018), we compare the objective value error of the path algorithm, ADMM, and the Minkowski methods relative to the final objective value of the Gurobi-solved quadratic program. Results for the zero-sum constrained lasso are shown in Figure F.2 for problem sizes $(n, d) = (100, 500)$, $(500, 1000)$, $(1000, 2000)$, $(2000, 4000)$, and $(4000, 8000)$ for which all four methods could be terminated. Likewise, results for the nonnegative lasso are presented in Figure F.3 for $(n, d) = (100, 500)$, $(500, 1000)$, and $(1000, 2000)$.

The presented results are consistent with those reported by Gaines et al. (2018) in case of the zero-sum lasso: the path algorithm is more accurate than the first-order methods; the accuracy of ADMM descreases as λ increases. It is notable that the Minkowski method tends to be more accurate than ADMM and is not sensitive to the sparsity level λ. Nevertheless, the accuracy of both first-order methods were less than 0.0001%, and this amount of error will not be significant practically.

On the other hand, in the nonnegative lasso the path-following algorithm was not as accurate as the other methods for $(n, d) = (100, 500)$, and at most similar for the larger problem sizes. Except for this outlier, all methods maintained high accuracy of less than $10^{-5}\%$ of relative error; ADMM was not sensitive to λ in this example.

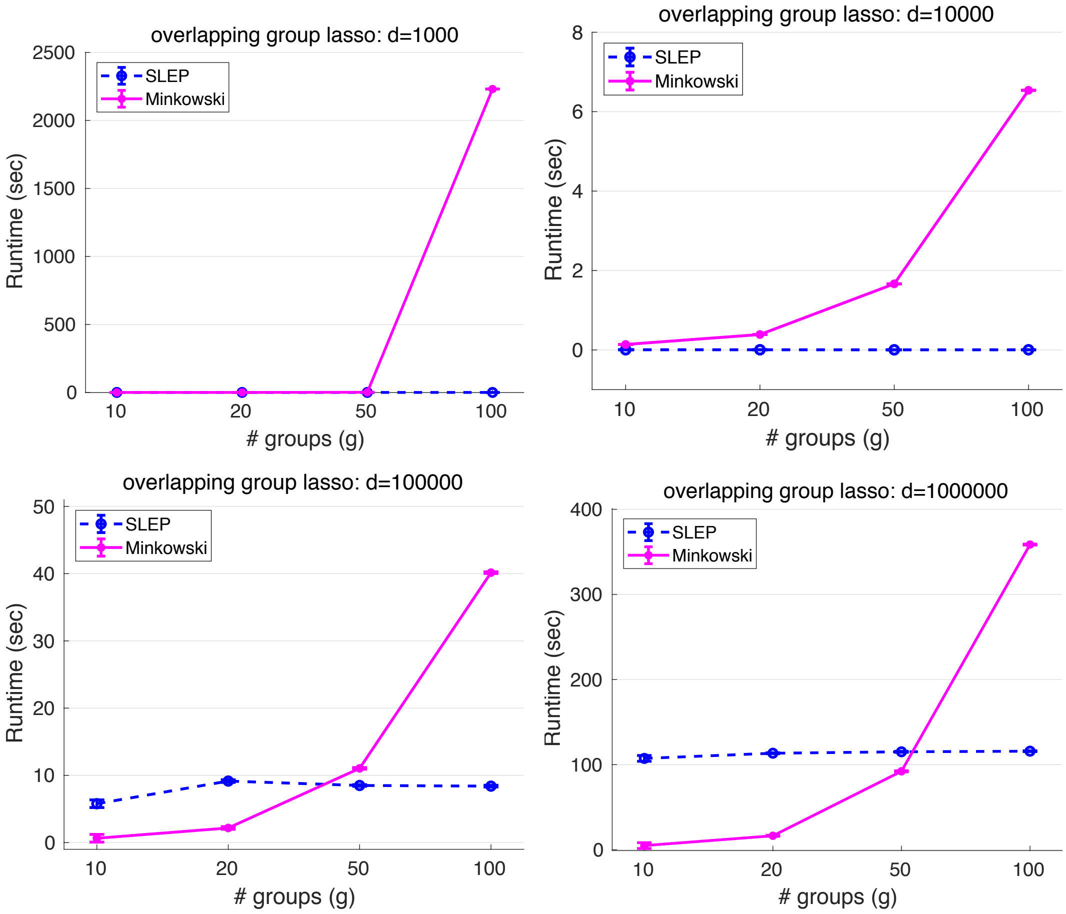

Figure E.2. Runtime-by-number-of-groups comparison of the proposed Minkowski method and the dual projected gradient method by Yuan et al. (2011) for overlapping group lasso.

References

Duan, L. L., Young, A. L., Nishimura, A., and Dunson, D. B. Bayesian constraint relaxation. *arXiv preprint arXiv:1801.01525*, 2018.

Gaines, B. R., Kim, J., and Zhou, H. Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4):861–871, 2018.

Lange, K. *MM Optimization Algorithms*. SIAM, 2016.

Patra, S. and Dunson, D. B. Constrained bayesian inference through posterior projections. *arXiv preprint arXiv:1812.05741*, 2018.

Xu, J., Chi, E., and Lange, K. Generalized linear model regression under distance-to-set penalties. In *Advances in Neural Information Processing Systems*, pp. 1385–1395, 2017.

Yuan, L., Liu, J., and Ye, J. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, pp. 352–360, 2011.

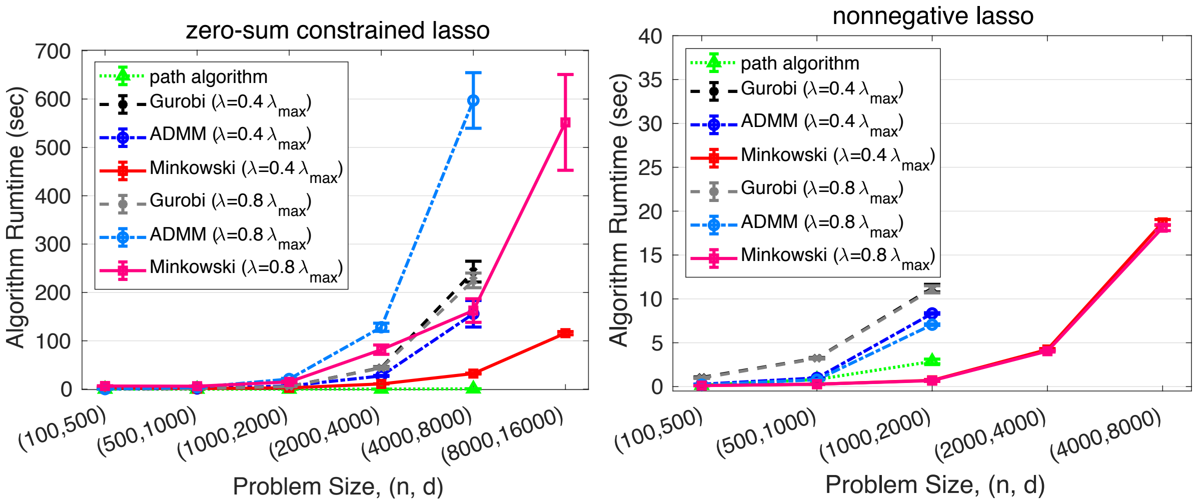

Figure F.1. Timing comparison of the proposed Minkowski method and the other methods by Gaines et al. (2018) for the constrained lasso. Left, runtime for the zero-sum constrained lasso. Right, runtime for the nonnegative lasso.

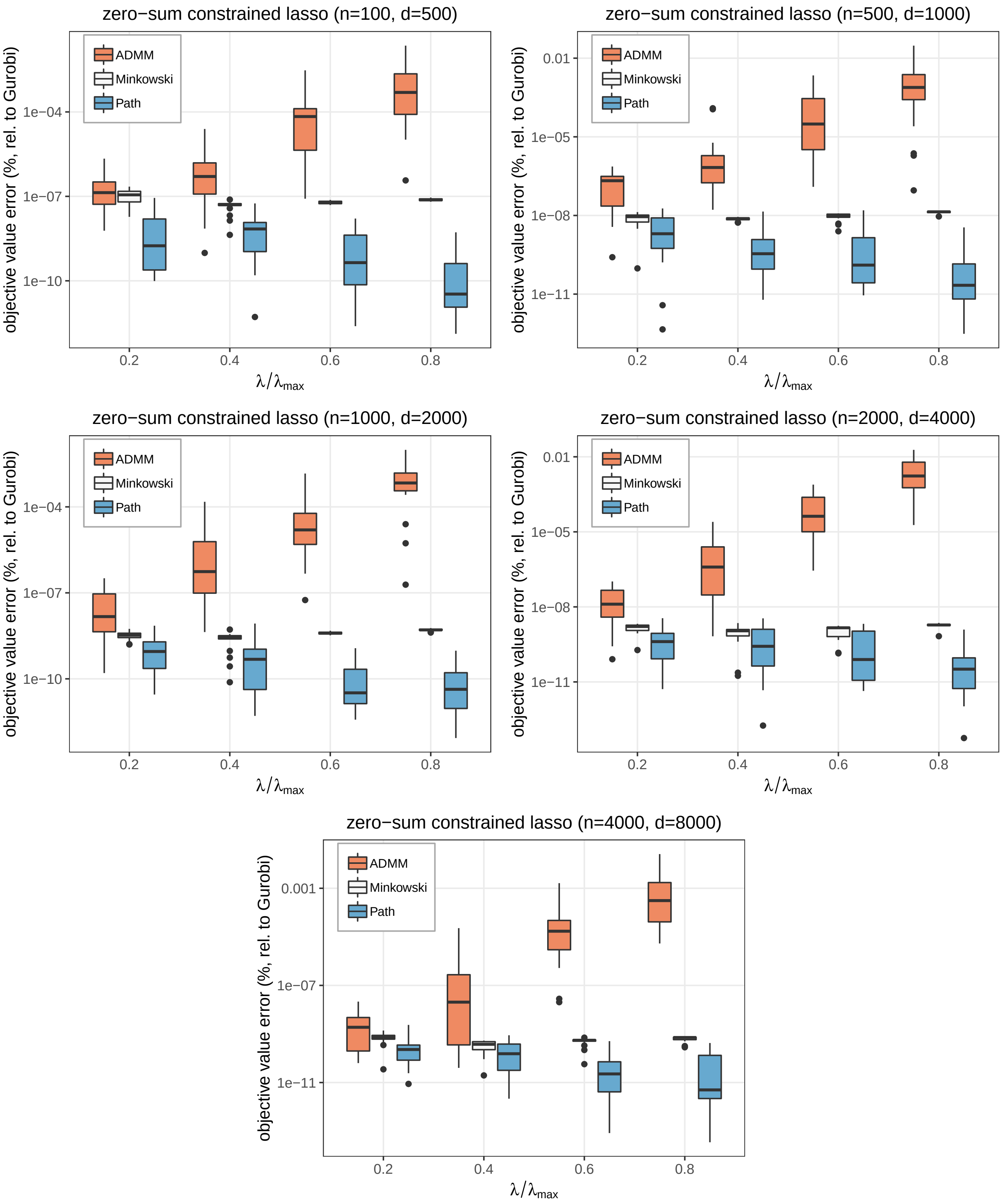

Figure F.2. Comparison of the proposed Minkowski method and the other methods by Gaines et al. (2018) for the zero-sum constrained lasso.

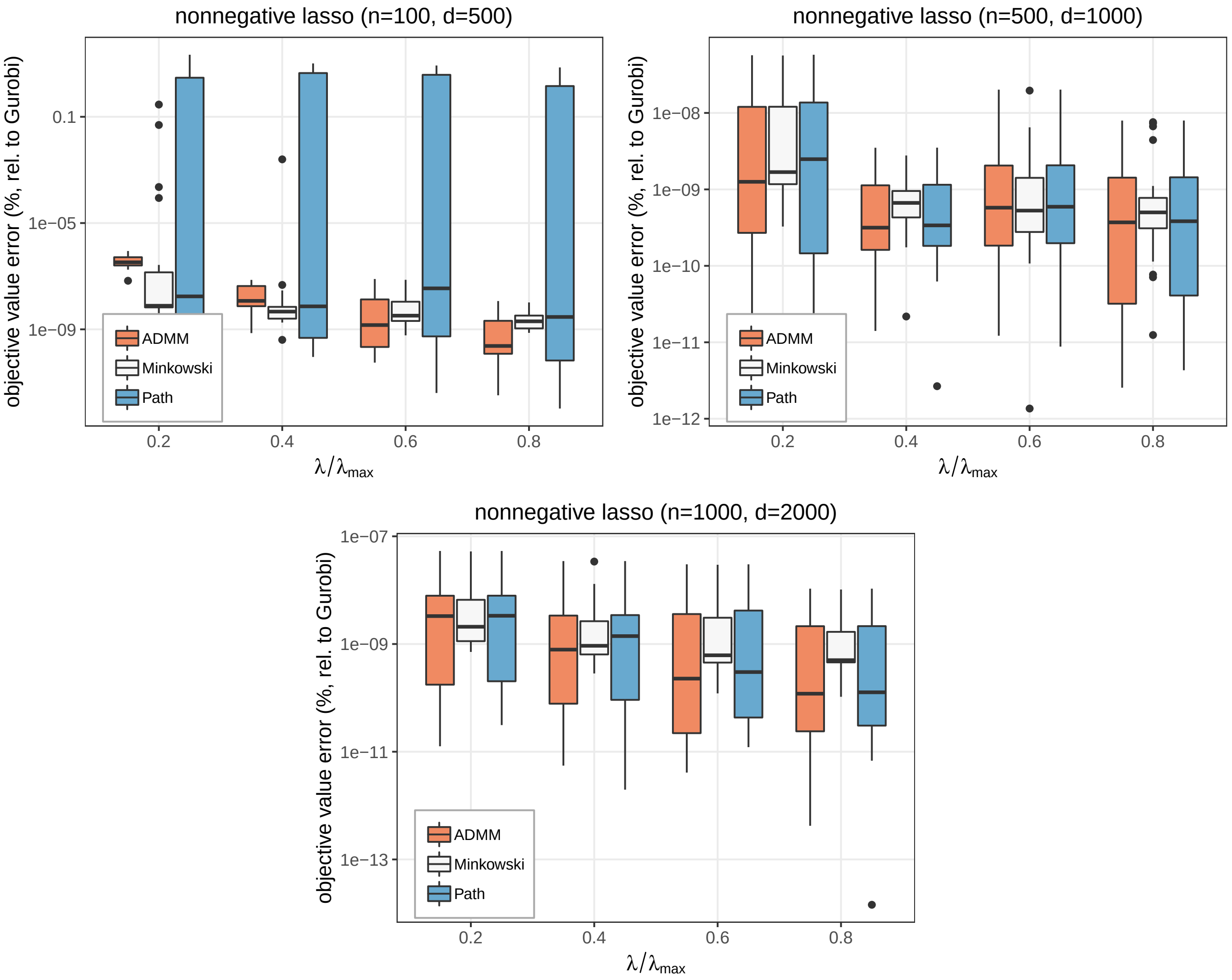

Figure F.3. Comparison of the proposed Minkowski method and the other methods by Gaines et al. (2018) for the nonnegative lasso.