We thank the reviewers for their thoughtful comments. To summarize, there is a consensus that the proximal analysis is theoretically sound and novel and that our method outperforms existing ones. Most concerns are regarding the presentation, scope, and technicality, not about the content. We believe concerns can be largely addressed by better exposition and targeted clarifications.

Presentation: We believe this submission is the first to notice the importance of the prox operator of the matrix perspective function ϕ. Refs [2, 19] only mention that ϕ is convex and stop there. We got interested in the prox of this function in need for constrained joint MLE of the two natural parameters of multivariate Gaussian (Eq 2). Most textbooks (e.g., [And09]) deal with fairly simple settings that can be solved analytically, or submit to suboptimal procedures. Proximal methods come to rescue for more complex settings at modern scales. Yet, computing the prox of ϕ turned out to be nontrivial, let alone doing it efficiently and accurately. We will add this point to §1 and reorganize it after merging §2 and §3, as suggested by R1. (Minor typos have already been fixed.)

Scope: Given the significance of the multivariate Gaussian, we think enlarging the class of solvable estimation problems is important and useful to the community, let alone other problems discussed in the submission. Joint estimation of Gaussian natural parameters under constraints have not received much attention, and we think a part of the reason is the lack of practical optimization algorithms. Please see Comments 3-3 by R2 and 2-2 by R3.

Technicality: We understand the concern on the high technicality of the submission. On the other hand, we think it is the nature of semismooth optimization. As the latter is starting to get applied to learning problems [LST18, CZST20], we also think this submission is timely to the community. We will add remarks on the meaning of the technical results as recommended by R1 and R4. They are roughly as follows. *Theorem 1*: computation of the prox operator of the nonsmooth function ϕ, which does not have a closed form, reduces to a univariate root-finding problem. *Theorem 2*: the function whose root is sought is nonsmooth, but strongly semismooth and satisfies conditions for a Newton algorithm to exist and converge locally. *Theorem 3*: convergence is global and the rate is asymptotically quadratic. *Theorem 4*: the subgradients that need to be computed for the semismooth Newton algorithm have a closed form.

R1: **1)** We will merge §2 and §3 to avoid redundancy. **2)** Proof of Thm 3: Yes, $0 \leq v_k \leq 1$ for all k. It follows from the Bolzano-Weierstrauss Theorem that $\{v_k\}$ has a limit point. Presentation will be improved once §2 and §3 are merged to give more space. **3)** Convergence result in Thm 3 is *global*, while asymptotic. The nonasymptotic analysis in [2] relies on strong convexity and Lipschitz continuity of the Hessian, neither of which does not apply to our problem. In many semismooth Newton literature [25, LST18, CZST20], the analysis is either local or asymptotic. (Asymptotically) quadratic convergence usually translates to that only a few iterations are required to get high accuracy, as can be seen in Table 1. Any (including stochastic) proximal algorithm will benefit from a fast, high-accuracy prox routine, since it is the former that is to run for a finite but large number of iterations.

R2: We think the motivation for the Newton method is delineated in the above "Presentation" paragraph. Bisection was ruled out since a) in a similar matrix nearness problem [3, §4] it was reported to be orders of magnitudes slower than Newton (when the latter is possible), and b) when we observed that the Newton algorithm took less than 10 iterations for all experiments with several orders of magnitudes better accuracy (in terms of the KKT measure) than the interior-point solver MOSEK, we thought that the game was over. We will add bisection to Table 1 and potentially convergence plot similar to [3, Fig 3] as suggested by R4.

R3: Comparison with [3] and [25] – Our dual (Eq 15) is similar to the "one rank-one" case of [3]. However, in [3, §3.4], that $\bar{\mathbf{X}} \succeq \mathbf{0}$ plays an important role in devising a (smooth) Newton algorithm; we do not assume this. Furthermore, quadratic convergence of the Newton algorithm is only claimed and not proved in [3]. In [25], the constraint is that all diagonal entries of $\mathbf{U}$ are 1. Rather surprisingly, our constraint that only one diagonal entry is 1 makes the perturbation analysis of the spectral decomposition of $\mathbf{C}(\mu)$ more difficult (Lemma A.3 and §A.4) than [25, Lemma 3.4].

R4: **1)** None of the three learning tasks illustrated in the submission possesses an obvious alternative solution method like FISTA or non-linear conjugate gradient acceleration of ISTA. ISTA-type methods require the objective to be the sum of a smooth and nonsmooth functions, where the former has a Lipschitzian gradient. Foremost, function ϕ is nonsmooth. Thus the *heteroskedastic* scaled lasso (Eq 4), unlike the plain lasso, cannot benefit from ISTA. In this case, a popular method is Chambolle-Pock [5], a special case of PDHG (see L47-48 and references therein). Variance-constrained joint Gaussian MLE (Eq 2) and pseudolikelihood graphical model selection (Eq 3) are *not* the same as the graphical lasso. The differentiable parts of the objectives do not have Lipschitzian gradients due to the logarithmic terms, hence PDHG was considered. Even in the graphical lasso, if the location parameter ($\boldsymbol{\eta}$) is to be jointly estimated, then the log-likelihood (+ ℓ_1 penalty on $\boldsymbol{\Omega}$) still contains the ϕ. **2)** For the convergence plot, see our response to R2.

[And09] Anderson, T. W. An Introduction to Multivariate Statistical Analysis. Wiley, 2009.
[CZST20] Chu, D. et al., "Semismooth Newton Algorithm for Efficient Projections onto $\ell_{1,\infty}$-norm Ball." ICML 2020.
[LST18] Li, X., Sun, D. and Toh, K.C. "A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems." SIAM Journal on Optimization, 28(1), pp.433-458, 2018.