We sincerely thank all the reviewers for their detailed comments and queries, and give clarifications and answers below.

Reviewer 1: We will revise according to all the comments on typos, clarity, and rigor of the mathematical writing. First, we focus on the correctness of Lemma 1, and then address the others comments.

Correctness. We agree with the reviewer that Lemma 1 is not precise, as we (mistakenly) did not include the precise range of x in the statements. We provide here the precise version of Lem.1, where the differences are colored in BLUE:

Lemma 1. *For a closed convex set $\mathbb{X}$, a convex proper l.s.c. function $f : \mathbb{X} \to \mathbb{R} \cup \{\infty\}$ and $\lambda > 0$ define $f_\lambda : \mathbb{R}^d \to \mathbb{R}$ as $f_\lambda(x) := \min_{x' \in \mathbb{X}} f(x') + \frac{1}{2\lambda}\|x - x'\|^2$ and $\hat{x}_\lambda(x) := \operatorname{argmin}_{x' \in \mathbb{X}} f(x') + \frac{1}{2\lambda}\|x - x'\|^2$. Then for any $x \in \mathbb{X}$:*
(a) $\hat{x}_\lambda(x)$ is unique and $f(\hat{x}_\lambda(x)) \le f_\lambda(x) \le f(x)$.
(b) f_λ is convex, differentiable, $1/\lambda$-smooth and $\nabla f_\lambda(x) = (1/\lambda)(x - \hat{x}_\lambda(x))$, and,
(c) if f is G-Lipschitz continuous, then, $\|\hat{x}_\lambda(x) - x\| \le G\lambda$, and $f(x) \le f_\lambda(x) + G^2\lambda/2$.

This version of Lemma 1 is (i) sufficient for proving our main results (lines 577, 622, 691, 705) and (ii) correct. Since the reviewer's counter example uses $x \notin \mathbb{X}$, it does not contradict Lemma 1(c). We now provide a full proof below.
Proof (brief due to page limit). Denote $\phi_{\lambda,x}(x') := f(x') + (1/2\lambda)\|x - x'\|^2$. Note that $\phi_{\lambda,x}(\cdot)$ is a $1/\lambda$-strongly convex function and $f_\lambda(x) = \min_{x' \in \mathbb{X}} \phi_{\lambda,x}(x')$.
(a) Then $f(\hat{x}_\lambda(x)) \le \phi_{\lambda,x}(\hat{x}_\lambda(x)) = \min_{x' \in \mathbb{X}} \phi_{\lambda,x}(x') = f_\lambda(x) \le \phi_{\lambda,x}(x) = f(x)$ and the uniqueness of $\hat{x}_\lambda(x)$ follows from the strong convexity of $\phi_{\lambda,x}(\cdot)$ and the fact that f is a proper convex function.
(b) Let $x \in \mathbb{R}^d$ and $g_x := (x - \hat{x}_\lambda(x))/\lambda$. By $1/\lambda$ strong convexity of $\phi_{\lambda,x}(x')$ and $\hat{x}_\lambda(x) = \operatorname{argmin}_{x' \in \mathbb{X}} \phi_{\lambda,x}(x')$, we have for any $x' \in \mathbb{X}$ that $\phi_{\lambda,x}(x') \ge \phi_{\lambda,x}(\hat{x}_\lambda(x)) + \|x' - \hat{x}_\lambda(x)\|^2/2\lambda$, which simplifies to $f(x') \ge f(\hat{x}_\lambda(x)) + \langle g_x, x' - \hat{x}_\lambda(x)\rangle$. Using this, for any $x, y \in \mathbb{R}^d$ we get

$$\begin{aligned} f_\lambda(y) - f_\lambda(x) &= f(\hat{x}_\lambda(y)) - f(\hat{x}_\lambda(x)) + (\|\hat{x}_\lambda(y) - y\|^2 - \|\hat{x}_\lambda(x) - x\|^2)/2\lambda \\ &\ge \langle g_x, \hat{x}_\lambda(y) - \hat{x}_\lambda(x)\rangle + \lambda/2(\|g_y\|^2 - \|g_x\|^2) = \langle g_x, y - x\rangle + \lambda/2\|g_x - g_y\|^2 \end{aligned} \tag{1}$$

Instantiating the above for $y \leftarrow x$, $x \leftarrow y$ we also get $f_\lambda(y) - f_\lambda(x) \le \langle g_y, y - x\rangle - \lambda/2\|g_x - g_y\|^2$. Combining these two inequalities

$$0 \le \lambda/2\|g_y - g_x\|^2 \le f_\lambda(y) - f_\lambda(x) - \langle g_x, y - x\rangle \le -\lambda/2\|g_y - g_x\|^2 + \langle g_y - g_x, y - x\rangle \le \|y - x\|^2/2\lambda \tag{2}$$

This implies that $\lim_{y \to x}(f_\lambda(y) - f_\lambda(x) - \langle g_x, y - x\rangle)/\|y - x\| = 0$. Thus f_λ is Frechet differentiable with gradient $\nabla f_\lambda(x) = g_x = (x - \hat{x}_\lambda(x))/\lambda$. The above inequality also implies f_λ is convex and $1/\lambda$-smooth.
(c) Let $x \in \mathbb{X}$. Using $1/\lambda$-strong convexity of $\phi_{\lambda,x}$ and $\hat{x}_\lambda(x) \in \operatorname{argmin}_{x' \in \mathbb{X}} \phi_{\lambda,x}(x')$, and G-Lipschitzness of f,

$$\begin{aligned} \|x - \hat{x}_\lambda(x)\|^2/2\lambda \le \phi_{\lambda,x}(x) - \phi_{\lambda,x}(\hat{x}_\lambda(x)) = f(x) - f_\lambda(x) &= f(x) - f(\hat{x}_\lambda(x)) - \|x - \hat{x}_\lambda(x)\|^2/2\lambda \\ &\le G\|\hat{x}_\lambda(x) - x\| - \|x - \hat{x}_\lambda(x)\|^2/2\lambda \le G^2\lambda/2 . \quad \square \end{aligned}$$

We say line 557: "for simplicity...$\mathbb{X}$ is the whole vector space". This was an assumption made, in the context of Sec. A.1, for ease of exposition of the failed attempt at a PO efficient algorithm (Algo. APGD).

Experimental verification. As suggested, we compared the projection-free methods using a higher-dimensional ($d = 50,176$) ImageNet dataset in the same low-rank SVM problem. For achieving an optimality gap of 0.02, Randomized-FW[52] used 34717 / 264 FO / LMO calls and our MOLES used 4004 / 241 FO / LMO calls. We will add detailed simulation results including sensitivity analysis in the next revision. We agree that our algorithms have more parameters and hence harder to tune than most baselines. Overcoming this is an important direction of future research.

Reviewer 2 We agree that the reviewer's definition of the stochastic subgradient oracle is more appropriate. We modified the manuscript according to the additional comments.

Reviewer 3 The two properties we need of the superset $\mathbb{X} \supseteq \mathcal{X}$ are that (a) it is easy to project onto $\mathbb{X}$ and (b) f is G-Lipschitz on $\mathbb{X}$. In our paper, we choose $\mathbb{X}$ to be a Euclidean ball (which is easy to project to) but any other choice of $\mathbb{X}$ which satisfies the above properties works just as well. One choice for this Euclidean ball is $B(x_0, D_{\mathcal{X}})$, where x_0 is the initial point and $D_{\mathcal{X}}$ is the diameter of $\mathcal{X}$, instead of the ball of radius $2R$ we currently use.

As mentioned by R3, even if f is G-Lipschitz inside the constraint $\mathcal{X}$, it could (i) blow up or (ii) be undefined just outside of $\mathcal{X}$. Thus an $\mathbb{X}$ satisfying our requirements may not exist. In our experiments, we do not explicitly project onto $\mathbb{X}$ (`line 3.16`) but still observed that $\|x_k - x'_k\| = O(G\lambda)$ and small, which implies that the iterates x'_k are close to $\mathcal{X}$. This hints that we may only need Lipschitzness over a much smaller set $\mathcal{X} + B(0, O(G\lambda))$, but we do know how to prove this yet. Theoretically, we can work around the issue (ii) above by minimizing the convex extension $f_{\mathcal{X}} : \mathbb{R}^d \to \mathbb{R}$ of the function f from the set $\mathcal{X}$, defined as $f_{\mathcal{X}}(x') := \max_{x \in \mathcal{X}} \max_{g \in \partial f(x)} f(x) + \langle g, x' - x\rangle$. The extension $f_{\mathcal{X}}$ has the same value as f inside $\mathcal{X}$ and is G-Lipschitz everywhere. Therefore the following minimization problems are equivalent: $\min_{x \in \mathcal{X}} f(x)$ and $\min_{x \in \mathcal{X}} f_{\mathcal{X}}(x')$. However, it is not clear if we can even estimate/approximate the gradients of $f_{\mathcal{X}}$ efficiently. We could not find any relevant prior work and leave this question for future work. We modified the manuscript according to the additional comments.