BEYOND POLYTOPES: DAMPED NEWTON FRANK–WOLFE METHODS OVER COMPACT CONVEX SETS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study convex optimization over compact convex sets for twice differentiable objective functions with positive definite Hessians, assuming access to the feasible set through a linear minimization oracle. Existing second-order Frank–Wolfe methods either provide only a local linear rate over general convex sets or achieve global linear and local quadratic convergence only over polytopes. We propose *damped Newton Frank–Wolfe methods* that approximately solve constrained damped Newton subproblems using Frank–Wolfe or away-step Frank–Wolfe inner iterations. We develop residual backtracking, using a root Newton stepsize as a safe reference, together with a switching rule that eventually takes full steps. Under Hölder smoothness assumptions for pth derivatives, we establish global linear convergence and local convergence with Q-order at least $1 + \nu$ ($p = 2$) and local quadratic convergence ($p = 3$). Experiments on matrix sensing and ridge-regularized logistic regression show that the residual-backtracking variant is consistently robust and often achieves the best performance among the tested first- and second-order baselines.

1 INTRODUCTION

In this paper, we focus on the constrained convex optimization problem of the form

$$\min_{x \in \mathcal{X}} \quad f(x) \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable convex function whose Hessian $\nabla^2 f(x)$ is positive definite for all $x \in \mathcal{X}$, and $\mathcal{X} \subset \mathbb{R}^n$ is a compact convex set. Let $x^\star$ denote a minimizer of (1.1). We assume access to a linear minimization oracle (LMO) for $\mathcal{X}$, *i.e.*, given $c \in \mathbb{R}^n$, we can compute $v \in \operatorname{argmin}_{u \in \mathcal{X}} \langle c, u \rangle$. The *Frank–Wolfe method* (FW) is a projection-free method that requires only LMO calls to solve (1.1) rather than projection oracles. A projection oracle can be computationally expensive for complex constraint sets, whereas an LMO is often more efficient as shown in (Combettes & Pokutta, 2021; Braun et al., 2025; Woodstock, 2026). From this fact, the FW method and LMO have been widely used in machine learning optimizers (Sfyraki & Wang, 2026) and signal processing (Odor et al., 2016; Takahashi et al., 2026a) as well as in optimization (Braun et al., 2025).

Related Work. The FW method was originally proposed by Frank & Wolfe (1956) and was independently developed and extended as the *conditional gradient method* (CG) by Levitin & Polyak (1966). The first lower bounds on the convergence rate of the FW method were established by Canon & Cullum (1968) and later refined by Guélat & Marcotte (1986). Subsequent work sharpened the convergence analysis and clarified the limitations of FW methods. In particular, Jaggi (2013) refined the convergence theory and derived a lower bound relating iterate sparsity to solution accuracy. A closely related lower bound was obtained by Lan (2013) through an oracle-complexity analysis of methods based on linear optimization. Faster convergence rates can be obtained under additional geometric assumptions on the feasible set. Both a faster sublinear rate and a linear convergence rate over strongly convex sets were derived by Garber & Hazan (2015), and the analysis was extended to uniformly convex sets by Kerdreux et al. (2021). On polytopes, however, the classical FW method can zigzag near an optimal face, leading to slow convergence. The *away-step FW method* was proposed by Wolfe (1970) to mitigate zigzagging, and its linear convergence was established by Lacoste-Julien & Jaggi (2015). Pedregosa et al. (2020) proposed a backtracking line-search strat-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

egy, and Takahashi et al. (2026a) extended this strategy to the Bregman setting. For a comprehensive review of FW methods, we refer the interested reader to the survey by Braun et al. (2025).

Limitations of Existing Methods. Even for strongly convex objective functions, the classical FW method can exhibit only sublinear convergence without additional geometric assumptions on $\mathcal{X}$. Thus, unlike projected gradient methods, it does not generally admit a linear convergence guarantee.

The *conditional gradient sliding method* (CGS) was proposed by Lan & Zhou (2016) to reduce the number of gradient evaluations for smooth convex problems while retaining the optimal LMO complexity. It incorporates Nesterov's acceleration and approximately solves the projection subproblems using multiple LMO calls, allowing each gradient evaluation to be reused across several inner iterations (see Section A for acceleration and projection methods). For strongly convex objectives, CGS achieves a linear rate in terms of gradient evaluations without additional geometric assumptions on $\mathcal{X}$. More recently, line-search-free adaptive stepsizes were incorporated into CGS by Takahashi (2026), removing the need for prior knowledge of a global smoothness constant.

Second-order FW methods have also been developed to obtain faster convergence. The *Newton FW method* (NFW), proposed by Liu et al. (2022), uses second-order information to achieve local linear convergence for constrained self-concordant problems (see Section A for second-order methods). NFW is globally convergent, but its linear rate is guaranteed only in the full-step phase. The *second-order conditional gradient sliding method* (SOCGS), proposed by Carderera & Pokutta (2026), approximately solves constrained Newton subproblems using an away-step FW method and achieves global linear convergence and local quadratic convergence for strongly convex objectives over polytopes. These guarantees are restricted to polytopes, with strict complementarity required for the local quadratic rate. Setting the subproblem accuracy also requires a lower bound on the primal gap $f(x_k) - f(x^\star)$ at the kth iteration. Then the following question naturally arises:

> *Can a projection-free Newton method combine global linear convergence with local quadratic convergence over compact convex sets?*

Contributions. We answer the above question in the affirmative by developing new Newton FW methods for general compact convex sets. Our main contributions are summarized as follows.

- **Algorithmic framework.** We propose a damped Newton FW method (Algorithm 1) and a variant with an away-step FW inner loop (Algorithm 3). Unlike existing second-order FW methods (Liu et al., 2022; Carderera & Pokutta, 2026), our methods are based on damped constrained Newton subproblems.
- **Stepsize and switching strategies.** We develop residual backtracking for the damped Newton FW method. Since the underlying Newton subproblems are constrained, their design and analysis differ from those for unconstrained Newton methods. We also introduce a switching rule that eventually sets $\alpha_k = 1$, thereby enabling local quadratic convergence.
- **Convergence guarantees.** We establish global linear convergence over compact convex sets without additional geometric assumptions. Switching yields local convergence with Q-order at least $1+\nu$ when $p = 2$ and $\nu > 0$, and local quadratic convergence when $p = 3$. Moreover, the inner accuracy does not require a lower bound on the primal gap.
- **Numerical performance.** Our method is consistently robust and often outperforms the tested first- and second-order baselines.

Table 1 compares the convergence rates of the proposed method with those of existing FW methods. Table 2 in Section D.4 shows the comparison of the total LMO complexity.

Notation. We use $\|\cdot\|$ to denote the Euclidean norm and $\langle\cdot,\cdot\rangle$ to denote the Euclidean inner product for vectors and Frobenius inner product for matrices. For a positive definite matrix H, define $\|y\|_H := \langle Hy, y\rangle^{1/2}$ and $\|h\|_H^* := \langle h, H^{-1}h\rangle^{1/2}$. We abbreviate $\|y\|_x := \|y\|_{\nabla^2 f(x)}$ and $\|h\|_x^* := \|h\|_{\nabla^2 f(x)}^*$, and define $D := \sup_{x,y\in\mathcal{X}} \|x - y\|_x$. Given $x \in \mathcal{X}$ (usually we set $x = x_k$), the operator norm of a matrix $H \in \mathbb{R}^{n\times n}$ induced by the Euclidean norm, and the operator norms of H and a third-order tensor $T \in \mathbb{R}^{n\times n\times n}$ are defined as

$$\|H\|_2 := \sup_{y\neq 0} \frac{\|Hy\|}{\|y\|}, \qquad \|H\|_{\mathrm{op}} := \sup_{y\neq 0} \frac{\|Hy\|_x^*}{\|y\|_x}, \qquad \|T\|_{\mathrm{op}} := \sup_{y,z,w\neq 0} \frac{|T[y,z,w]|}{\|y\|_x\|z\|_x\|w\|_x},$$

Table 1: Global and local complexities (measured in outer iterations) of LMO-based methods for achieving $f(x_k) - f(x^\star) \le \epsilon$. r-CGS denotes restarted CGS. SOCGS requires strict complementarity for its local quadratic rate. Our $\mathcal{C}^2$ bounds require $\nu > 0$. Total LMO complexities are reported in Table 2 (Section D.4).

Algorithm	f	∇f Lip.	$\nabla^p f$	$\mathcal{X}$	global	local
FW	$\mathcal{C}^1$ cvx	✔	✘	cvx	$\mathcal{O}(\epsilon^{-1})$	—
CGS	$\mathcal{C}^1$ cvx	✔	✘	cvx	$\mathcal{O}(\epsilon^{-1/2})$	—
r-CGS	$\mathcal{C}^1$ scvx	✔	✘	cvx	$\mathcal{O}(\log \epsilon^{-1})$	—
NFW	$\mathcal{C}^3$ cvx	✘	self-conc.	cvx	—	$\mathcal{O}(\log \epsilon^{-1})$
SOCGS	$\mathcal{C}^2$ scvx	✔	$\nabla^2 f$ Lip.	polytope	$\mathcal{O}(\log \epsilon^{-1})$	$\mathcal{O}(\log\log \epsilon^{-1})$
Alg. 1, 3	$\mathcal{C}^2$ scvx	✔	$\nabla^2 f$ Hölder	cvx	$\mathcal{O}(\log \epsilon^{-1})$	$\mathcal{O}(\log\log \epsilon^{-1})$
Alg. 1, 3	$\mathcal{C}^3$ scvx	✔	$\nabla^3 f$ Hölder	cvx	$\mathcal{O}(\log \epsilon^{-1})$	$\mathcal{O}(\log\log \epsilon^{-1})$

respectively. We define the FW gap at x_k by $g_k := \max_{v\in\mathcal{X}} \langle \nabla f(x_k), x_k - v\rangle$. For a set $P \subset \mathbb{R}^n$, $\operatorname{conv} P$ and $\operatorname{Vert} P$ denote its convex hull and set of extreme points, respectively.

2 Proposed Method: Damped Newton Frank–Wolfe Method

We propose projection-free methods for solving (1.1) based on second-order information of f.

Damped Newton FW Method. We define the damped quadratic model of f at x_k as

$$\Psi_\alpha(x; x_k) := f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2\alpha}\|x - x_k\|_{x_k}^2.$$

When $\alpha = 1$, $\Psi_\alpha(\cdot; x_k)$ coincides with the second-order Taylor model of f at x_k. For $\alpha \in (0, 1)$, the coefficient of the quadratic term is increased, and hence deviations from x_k are penalized more strongly. Thus, α serves as a stepsize, or equivalently, as a damping parameter for the Newton step. Indeed, minimizing $\Psi_\alpha(\cdot; x_k)$ over $\mathcal{X}$ is equivalent to computing the scaled projection

$$\operatorname*{argmin}_{x\in\mathcal{X}} \Psi_\alpha(x; x_k) = \operatorname*{argmin}_{x\in\mathcal{X}} \left\|x - \left(x_k - \alpha\nabla^2 f(x_k)^{-1}\nabla f(x_k)\right)\right\|_{x_k}^2. \tag{2.1}$$

Consequently, for $\mathcal{X} = \mathbb{R}^n$, the unique minimizer is $x_k - \alpha\nabla^2 f(x_k)^{-1}\nabla f(x_k)$. For constrained problems, minimizing $\Psi_\alpha(\cdot; x_k)$ over $\mathcal{X}$ yields a projected variable-metric step (Carderera & Pokutta, 2026); projected Newton methods under simple constraints were studied in (Bertsekas, 1982). This fits the inexact proximal Newton method via an indicator function (Lee et al., 2014); our methods achieve global linear and fast local convergence over compact convex sets (see Section 3).

Minimizing $\Psi_\alpha(\cdot; x_k)$ requires solving a constrained quadratic problem, which may be computationally expensive. We therefore approximately minimize $\Psi_\alpha(\cdot; x_k)$ using FW, which accesses $\mathcal{X}$ only through LMO calls. Newton FW methods based on the undamped $\Psi_1(\cdot; x_k)$ have been proposed independently (Liu et al., 2022; Carderera & Pokutta, 2026). Liu et al. (2022) provide only a local linear rate, whereas the convergence guarantees of Carderera & Pokutta (2026) are restricted to polytopes.

Algorithm 1: Damped Newton FW Method

Input: $x_0 \in \mathcal{X}$, $\alpha_k \in (0, 1]$, $\eta_k \ge 0$
for $k = 0, 1, 2, \ldots$ **do**
  $x_{k+1} \leftarrow \texttt{FW}(\nabla f(x_k), \frac{1}{\alpha_k}\nabla^2 f(x_k), x_k, \eta_k)$

Algorithm 2: $\texttt{FW}(h, H, u, \eta)$

$u_0 \leftarrow u$
for $t = 0, 1, 2, \ldots$ **do**
  $v_t \leftarrow \operatorname{argmin}_{v\in\mathcal{X}} \langle h + H(u_t - u), v\rangle$
  if $\langle h + H(u_t - u), u_t - v_t\rangle \le \eta$ **then**
    return u_t
  $\gamma_t \leftarrow \min\left\{1, \frac{\langle h + H(u_t - u), u_t - v_t\rangle}{\|v_t - u_t\|_H^2}\right\}$
  $u_{t+1} \leftarrow (1 - \gamma_t)u_t + \gamma_t v_t$

In this paper, we consider the damped quadratic model $\Psi_\alpha(\cdot; x_k)$ with $\alpha \in (0, 1]$. We now present the *damped Newton FW method* (Algorithm 1), which approximately minimizes $\Psi_\alpha(\cdot; x_k)$ over $\mathcal{X}$ using FW (Algorithm 2). We terminate Algorithm 2 when the FW gap of $\Psi_\alpha(\cdot; x_k)$ is sufficiently small, *i.e.*, when it holds that

$$G_t := \max_{v\in\mathcal{X}} \langle \nabla\Psi_{\alpha_k}(u_t; x_k), u_t - v\rangle \le \eta_k. \tag{2.2}$$

We distinguish the model FW gap G_t at the tth inner iteration from the FW gap g_k of f at the kth iteration. Algorithm 2 is efficient because it uses the exact stepsize $\gamma_t = \operatorname{argmin}_{\gamma\in[0,1]} \Psi_{\alpha_k}((1-\gamma)u_t + \gamma v_t; x_k)$ with $h = \nabla f(x_k)$, $H = \frac{1}{\alpha_k}\nabla^2 f(x_k)$, and $u = x_k$. The choices of α_k and η_k in Section 3 guarantee convergence of the outer iterates. We analyze the complexity of the inner FW iterations in Algorithm 2 up to an accuracy η_k. The proof is given in Section B.1.

Proposition 2.1. Let $\alpha_k \in (0, 1]$ and $\eta_k > 0$ be given. For any $k \geq 0$, the number of inner iterations at the kth outer iteration is at most $\lceil 24D^2/(\alpha_k\eta_k)\rceil$.

Away-Step FW Inner Loop. We may also use away-step FW for (2.1). For this variant, we assume that the LMO returns a vertex of $\mathcal{X}$, including at initialization. It maintains a finite active set $\mathcal{S}_t \subset \operatorname{Vert}\mathcal{X}$ with $u_t \in \operatorname{conv}\mathcal{S}_t$ and supplements standard FW steps toward a vertex with away steps that can remove unfavorable active vertices. On polytopes, these steps can mitigate zigzagging near the optimal face (Wolfe, 1970; Guélat & Marcotte, 1986). For strongly convex L-smooth objectives over polytopes, it converges linearly (Lacoste-Julien & Jaggi, 2015). At each outer iteration, SOCGS (Carderera & Pokutta, 2026) computes an away-step FW update on f and an approximate solution of (2.1) with $\alpha_k \equiv 1$, then selects the point with lower objective value. Our method uses away-step FW only to approximately solve the subproblem, avoiding the extra update on f and candidate comparison while preserving the outer convergence analysis.

We present the damped Newton FW method with an away-step FW inner loop in Algorithm 3 (see Algorithm 5 in Section B.2 for `AFW`).

Algorithm 3: Damped Newton FW Method with `AFW`

Input: $x_0 \in \operatorname{argmin}_{v\in\mathcal{X}}\langle\nabla f(x), v\rangle$ for $x \in \mathcal{X}$, $\alpha_k \in (0, 1]$, $\eta_k \geq 0$

$\widehat{\mathcal{S}}_0 \leftarrow \{x_0\}$, $\widehat{\lambda}_0 \leftarrow 1$

for $k = 0, 1, 2, \ldots$ **do**

  $(x_{k+1}, \widehat{\mathcal{S}}_{k+1}, \widehat{\lambda}_{k+1}) \leftarrow \texttt{AFW}(\nabla f(x_k), \frac{1}{\alpha_k}\nabla^2 f(x_k), x_k, \eta_k, \widehat{\mathcal{S}}_k, \widehat{\lambda}_k)$

When $\mathcal{X}$ is a polytope, we analyze the complexity of the inner iterations up to η_k. The dependence on the inner accuracy is logarithmic rather than inverse-linear. The proof is given in Section B.2.

Proposition 2.2. Let $\alpha_k \in (0, 1]$ and $\eta_k > 0$ be given and let $\mathcal{X}$ be a polytope that is not a singleton. For any $k \geq 0$, the number of inner iterations at the kth outer iteration is at most $|\widehat{\mathcal{S}}_k| - 1 + 2\left\lceil \frac{\log\max\{1, g_k/\zeta_k\}}{-\log\left(1-\mu_k\delta_{\mathcal{X}}^2/(4L_k d_{\mathcal{X}}^2)\right)}\right\rceil$, where $\delta_{\mathcal{X}}$ is the pyramidal width of $\mathcal{X}$, $d_{\mathcal{X}} \coloneqq \max_{x,y\in\mathcal{X}}\|x - y\|$, $\mu_k \coloneqq \lambda_{\min}(\nabla^2 f(x_k))/\alpha_k$, $L_k \coloneqq \lambda_{\max}(\nabla^2 f(x_k))/\alpha_k$, and $\zeta_k \coloneqq \min\left\{L_k d_{\mathcal{X}}^2/2, \eta_k^2/(2L_k d_{\mathcal{X}}^2)\right\}$.

3 Convergence Analysis

We now establish the global convergence and the local convergence of Algorithms 1 and 3.

3.1 Global Convergence

We first state Hölder continuity assumptions of the pth derivative of f used in the analysis.

Assumption 3.1. The objective function f is p times differentiable and its pth derivative is Hölder continuous, *i.e.*, for some $\nu \in [0, 1]$, there exists a constant $L_{p,\nu} > 0$ such that, for all $x, y \in \mathcal{X}$,

$$\|\nabla^p f(x) - \nabla^p f(y)\|_{\mathrm{op}} \leq L_{p,\nu}\|x - y\|_x^{\nu}. \tag{3.1}$$

When $p = 3$, suppose that $\Gamma_{3,\mathcal{X}} \coloneqq \sup_{x\in\mathcal{X}}\|\nabla^3 f(x)\|_{\mathrm{op}}$ is finite. Throughout the $p = 3$ analysis, $L_{3,\nu}$ denotes a constant chosen sufficiently large so that $\Gamma_{3,\mathcal{X}} \leq 2\left(L_{3,\nu}/(1+\nu)\right)^{1/(1+\nu)}$.

Let $p \in \{2, 3\}$, $q \coloneqq p + \nu$, $\rho \in (0, 1)$, and $\omega > 0$. We define $r_k \coloneqq \nabla f(x_k) + s_{k-1}$ and $w_k \coloneqq \nabla f(x_k) + s_k$, where $s_k \coloneqq -\nabla f(x_k) - \frac{1}{\alpha_k}\nabla^2 f(x_k)(x_{k+1} - x_k)$. We initialize $s_{-1} = 0$, $\eta_{-1} = 0$, and $\Delta_{-1} = 0$, and set $\Delta_k \coloneqq \max\{\|r_k\|_{x_k}^*, \rho\Delta_{k-1}\}$. If the FW gap $g_k = 0$, the method

terminates at an optimal solution; otherwise, $\Delta_k > 0$. The inner accuracy for global convergence is

$$\eta_k := \frac{\omega \Delta_k^2}{1 + B_q \Delta_k^{(q-2)/(q-1)}}, \tag{3.2}$$

where $B_q := (9L_{p,\nu})^{\frac{1}{q-1}} \kappa_\omega^{\frac{q-2}{q-1}}$ and $\kappa_\omega := \frac{1+\sqrt{1+4\omega(1+\rho^{-2})}}{2}$.

Root Newton Stepsize. Hanzely et al. (2026) proposed a root Newton stepsize for unconstrained problems. We consider $\alpha_k := \frac{1}{1+\theta_k}$ and $\theta_k := B_q \Delta_k^{(q-2)/(q-1)}$. As shown below, this value guarantees acceptance and therefore provides a safe endpoint for backtracking. We instead select θ_k by backtracking and use $\alpha_k = (1+\theta_k)^{-1}$, while keeping (3.2) fixed throughout the search.

Residual Backtracking. Hanzely et al. (2026) also proposed the universal backtracking strategy for unconstrained problems. We extend it to constrained problems with inexact updates and cap the trial damping parameter. We present the *residual backtracking strategy* in Algorithm 4. Starting with $\bar{\theta}_0 > 0$ and $\tau > 1$, it uses the procedure in Algorithm 4 at each outer iteration. The trial quantities $x_{k,j}$, $s_{k,j}$, $w_{k,j}$, and $r_{k+1,j}$ are computed as in the algorithm, and the first trial satisfying

$$\langle r_{k+1,j}, \nabla^2 f(x_k)^{-1} w_{k,j} \rangle \geq \frac{c_p}{\alpha_{k,j}\theta_{k,j}} \|r_{k+1,j}\|_{x_k}^{*2} \tag{3.3}$$

is accepted. Here, $c_p := 1/2$ for $p = 2$ and $c_p := 1/4$ for $p = 3$. The accepted θ_k need not equal the root value, and η_k is not recomputed from it.

Algorithm 4: Residual Backtracking for Damped Newton FW

Input: x_k, $\Delta_k > 0$, $\bar{\theta}_k > 0$, $\tau > 1$
1 Compute η_k using (3.2)
2 **for** $j = 0, 1, 2, \ldots$ **do**
3 $\quad \theta_{k,j} \leftarrow \min\{\tau^j \bar{\theta}_k, B_q \Delta_k^{(q-2)/(q-1)}\}$, $\alpha_{k,j} \leftarrow 1/(1+\theta_{k,j})$
4 $\quad x_{k,j} \leftarrow \texttt{FW}(\nabla f(x_k), \alpha_{k,j}^{-1} \nabla^2 f(x_k), x_k, \eta_k)$
5 $\quad s_{k,j} \leftarrow -\nabla f(x_k) - \alpha_{k,j}^{-1} \nabla^2 f(x_k)(x_{k,j} - x_k)$
6 $\quad w_{k,j} \leftarrow \nabla f(x_k) + s_{k,j}$, $r_{k+1,j} \leftarrow \nabla f(x_{k,j}) + s_{k,j}$
7 $\quad$ **if** $\langle r_{k+1,j}, \nabla^2 f(x_k)^{-1} w_{k,j} \rangle \geq \frac{c_p}{\alpha_{k,j}\theta_{k,j}} \|r_{k+1,j}\|_{x_k}^{*2}$ **then**
8 $\quad\quad \theta_k \leftarrow \theta_{k,j}$, $\alpha_k \leftarrow \alpha_{k,j}$, $x_{k+1} \leftarrow x_{k,j}$, $s_k \leftarrow s_{k,j}$, $\bar{\theta}_{k+1} \leftarrow \theta_k/\tau$
9 $\quad\quad$ **break**

Each call to `FW` may be replaced by `AFW` in Line 4, initialized with a copy of the current active-set representation. Only the active set and weights of the accepted trial are carried to the next outer iteration. The following residual estimate applies to both strategies and to each trial satisfying the inner stopping condition. Its proof is given in Section C.1.

Lemma 3.2. For every $k \geq 0$, the following inequality holds:

$$\|w_k\|_{x_k}^{*2} \leq \|r_k\|_{x_k}^* \|w_k\|_{x_k}^* + \frac{\eta_k + \eta_{k-1}}{\alpha_k}. \tag{3.4}$$

With (3.2) at each iteration, $\alpha_k \geq (1 + B_q \Delta_k^{(q-2)/(q-1)})^{-1}$ implies $\|w_k\|_{x_k}^* \leq \kappa_\omega \Delta_k$.

Residual backtracking satisfies the following one-step decrease; its proof is given in Section C.3.

Theorem 3.3 (Sufficient decrease)**.** Suppose that Assumption 3.1 holds and use (3.2). For $\Delta_k > 0$, let $x_{k+1} \in \mathcal{X}$ satisfy (2.2). If either $\theta_k = B_q \Delta_k^{(q-2)/(q-1)}$ or θ_k is accepted by Algorithm 4, then

$$f(x_k) - f(x_{k+1}) \geq \frac{c_p}{B_q} \frac{\|r_{k+1}\|_{x_k}^{*2}}{\Delta_k^{(q-2)/(q-1)}} - \eta_k. \tag{3.5}$$

For $w_{k,j} \neq 0$, Proposition C.1 in Section C.2 gives the following sufficient condition for (3.3):

$$\theta_{k,j} \geq (9L_{p,\nu})^{\frac{1}{q-1}} \|w_{k,j}\|_{x_k}^{*\frac{q-2}{q-1}}. \tag{3.6}$$

The root value guarantees acceptance by Lemma 3.2; the case $w_{k,j} = 0$ is immediate. Consequently, truncation guarantees finite backtracking, as stated below and proved in Section C.4.

Corollary 3.4 (Finite backtracking)**.** Suppose that Assumption 3.1 holds with $p \in \{2, 3\}$ and $\nu \in [0, 1]$. Let $\bar{\theta}_0 > 0$ and $\tau > 1$. With the accuracy (3.2), Algorithm 4 terminates its inner solves and backtracking after finitely many iterations at every nonterminal outer iteration.

Global Linear Convergence. Under Assumption 3.1 with $q > 2$, there exist $\mu, L > 0$ such that $\mu I \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathcal{X}$ (see Lemma C.5). We use the sufficient decrease just established to control the potential, without requiring residual backtracking to accept the root stepsize. The proof of the following proposition is given in Section C.5.

Proposition 3.5. Suppose that Assumption 3.1 holds with $q > 2$. Let $\omega > 0$ satisfy

$$0 < \omega < \frac{c_p \mu}{L} \rho^2 (1 - \rho^{q/(q-1)}). \tag{3.7}$$

Then, there exist constants $\chi, c_\chi > 0$ such that $\Phi(x_k) \coloneqq f(x_k) - f(x^\star) + \chi \Delta_k^{\frac{q}{q-1}}$ satisfies

$$\Phi(x_k) - \Phi(x_{k+1}) \geq c_\chi \frac{\Delta_{k+1}^2}{\Delta_k^{\frac{q-2}{q-1}}}. \tag{3.8}$$

Condition (3.7) ensures that $\chi > 0$ can be chosen in (C.17), with c_χ defined by (C.18). We now state the global convergence of Algorithms 1 and 3; its proof is given in Section C.6 (see also Section C for terminology).

Theorem 3.6 (Global R-linear convergence)**.** Suppose that Assumption 3.1 holds with $q > 2$ and that ω satisfies (3.7). Then, for all $k \geq 0$,

$$f(x_k) - f(x^\star) \leq \Phi(x_k) \leq \left(1 + \frac{c_\chi \rho^{(q-2)/(q-1)}}{\chi + (L/(2\mu) + \omega\rho^{-2}) \left(\Phi(x_0)/\chi\right)^{(q-2)/q}} \right)^{-k} \Phi(x_0). \tag{3.9}$$

3.2 Local Convergence

We now establish local convergence of Algorithms 1 and 3 with Q-order at least $1 + \nu$ for $p = 2$ and Q-quadratic convergence for $p = 3$. To combine global convergence with fast local convergence, we use the FW-gap-based condition below to switch to full steps after finitely many iterations, rather than merely relying on $\alpha_k \to 1$. Upon switching, we stop Algorithm 4, set $\alpha_k = 1$, and adopt the following inner stopping rule. When the FW gap $g_k > 0$, which implies $\|\nabla f(x_k)\| > 0$, define $M_t \coloneqq f(x_k) - \Psi_1(u_t; x_k)$. For any $a > 0$, define the stopping iteration by

$$\ell_k(a) \coloneqq \min \left\{ t \geq 0 \;\middle|\; G_t \leq \left(\frac{M_t}{\|\nabla f(x_k)\|} \right)^a \right\}. \tag{3.10}$$

This yields local convergence with Q-order at least $1 + \nu$ for both the iterates and the primal gap when $p = 2$, and local Q-quadratic convergence of both quantities when $p = 3$. In both cases, the resulting local iteration complexity is $\mathcal{O}(\log\log \epsilon^{-1})$. More precisely, when $p = 2$, the local iteration complexity is $\mathcal{O}(\log\log \epsilon^{-1} / \log(1 + \nu))$, which differs only by a constant factor for fixed $\nu > 0$. The following lemma establishes finite termination of the inner loop and bounds the error of its output. Its proof is given in Section D.1.

Lemma 3.7 (Finite termination and error estimate)**.** Suppose that $g_k > 0$ and $\alpha_k = 1$. Then, for any $a > 0$, the stopping iteration $\ell_k(a)$ in (3.10) is finite. Moreover, the following inequalities hold:

$$G_{\ell_k(a)} \leq \|\widehat{x}_{k+1} - x_k\|^a, \tag{3.11}$$

$$\|u_{\ell_k(a)} - \widehat{x}_{k+1}\| \leq \sqrt{\frac{2}{\mu}} \|\widehat{x}_{k+1} - x_k\|^{a/2}, \tag{3.12}$$

where $\widehat{x}_{k+1}$ is the exact minimizer of $\Psi_1(\cdot; x_k)$ over $\mathcal{X}$.

For $p = 2$, let K_2 be the switching iteration defined in (3.13). For $k < K_2$, we use Algorithm 4 with the accuracy (3.2). For $k \geq K_2$, if $g_k = 0$, the algorithm terminates; otherwise, we set $\alpha_k = 1$, run the inner method until (3.10) with $a = 2(1 + \nu)$ is satisfied, and set $x_{k+1} \coloneqq u_{\ell_k(2(1+\nu))}$. We now establish the local convergence of Algorithms 1 and 3 when $p = 2$ and $\nu \in (0, 1]$. Its proof is given in Section D.2.

Theorem 3.8 (Local convergence with Q-order at least $1+\nu$)**.** Suppose that Assumption 3.1 holds with $p=2$ and $\nu \in (0,1]$ and that ω satisfies (3.7). Let the switching iteration K_2 be defined as

$$K_2 := \min\left\{k \geq 0 \;\middle|\; g_k \leq \frac{\mu}{2}(2C_2)^{-\frac{2}{\nu}}\right\}, \tag{3.13}$$

where $C_2 := \frac{L_{2,\nu}L^{(1+\nu)/2}}{(1+\nu)\sqrt{\mu}} + \sqrt{\frac{2}{\mu}}\left(1+\frac{L_{2,\nu}L^{(1+\nu)/2}D^{\nu}}{(1+\nu)\mu^{(1+\nu)/2}}\right)^{1+\nu}$. Then, K_2 is finite, and either the sequence reaches $x^\star$ in finitely many iterations, or

$$\|x_{k+1}-x^\star\| \leq C_2\|x_k-x^\star\|^{1+\nu} \tag{3.14}$$

holds for every $k \geq K_2$. Moreover, in the latter case, there exists a constant $\widetilde{C}_2 > 0$ such that

$$f(x_{k+1})-f(x^\star) \leq \widetilde{C}_2\left(f(x_k)-f(x^\star)\right)^{1+\nu} \tag{3.15}$$

holds for all $k \geq K_2$, yielding convergence of both the iterates and primal gap with Q-order at least $1+\nu$ (Q-quadratic when $\nu=1$).

For $p=3$, Assumption 3.1 and compactness of $\mathcal{X}$ imply that, for some $M>0$,

$$\|\nabla^2 f(x)-\nabla^2 f(y)\|_2 \leq M\|x-y\|, \qquad x,y \in \mathcal{X}. \tag{3.16}$$

Let K_3 be the switching iteration defined in (3.17). For $k < K_3$, we likewise use Algorithm 4 with (3.2). For $k \geq K_3$, if $g_k = 0$, the algorithm terminates; otherwise, we set $\alpha_k = 1$, run the inner method until (3.10) with $a=4$ is satisfied, and set $x_{k+1} := u_{\ell_k(4)}$. We now establish the local convergence of Algorithms 1 and 3 when $p=3$ and $\nu \in [0,1]$. Its proof is given in Section D.3.

Theorem 3.9 (Local quadratic convergence)**.** Suppose that Assumption 3.1 holds with $p=3$ and $\nu \in [0,1]$ and that ω satisfies (3.7). Let the switching iteration K_3 be defined as

$$K_3 := \min\left\{k \geq 0 \;\middle|\; g_k \leq \frac{\mu}{8}C_3^{-2}\right\}, \tag{3.17}$$

where $C_3 := \frac{M}{2\mu} + \sqrt{\frac{2}{\mu}}\left(1+\frac{MD}{2\mu^{3/2}}\right)^2$. Then, K_3 is finite, and either the sequence reaches $x^\star$ in finitely many iterations, or

$$\|x_{k+1}-x^\star\| \leq C_3\|x_k-x^\star\|^2 \tag{3.18}$$

holds for every $k \geq K_3$. Moreover, in the latter case, there exists a constant $\widetilde{C}_3 > 0$ such that

$$f(x_{k+1})-f(x^\star) \leq \widetilde{C}_3\left(f(x_k)-f(x^\star)\right)^2 \tag{3.19}$$

holds for all $k \geq K_3$, yielding Q-quadratic convergence of both the iterates and primal gap.

4 Numerical Experiments

In this section, we present numerical experiments to demonstrate the performance of our proposed methods. All experiments were conducted in Julia 1.12 on a Mac Studio equipped with an Apple M4 Max processor and 128 GB of LPDDR5 memory. We compared Algorithms 1 and 3 using residual backtracking (RBNFW; see Algorithm 4) with the Newton FW method (NFW) (Liu et al., 2022), SOCGS (Carderera & Pokutta, 2026), the FW method with line search, the away-step FW method, CGS (Lan & Zhou, 2016), the projected gradient method (PG), and the accelerated projected gradient method (APG) (Nesterov, 1983; 2018; Beck & Teboulle, 2009). We applied SOCGS and the away-step FW method only to polytope settings. The descriptions in parentheses following RBNFW specify the definition of η_k. The setting labeled `global` used the accuracy (3.2) with $p=2$ for RBNFW. The maximal number of inner iterations was set to 1000 for RBNFW, NFW, CGS, and SOCGS. When an inner solve reached the cap before satisfying its prescribed accuracy condition, we accepted the last inner iterate as the next outer iterate (see Table 3 in Section E.3 for the number of outer iterations reaching this limit and its discussion). For the settings labeled `local2` and `local3`, we computed the constants in the switching criteria (3.13) and (3.17), respectively, and used (3.10) after switching, with $a = 2(1+\nu)$ and $a=4$, respectively. We set $\rho = 0.625$, $\bar{\theta}_0 = 0.25$, $\tau = 2$, and ω to 0.99 times the right-hand side of (3.7). In all experiments reported here, we set $\nu = 1$. All methods were set to terminate when the FW gap was below 10^{-8}. RBNFW, NFW, and SOCGS were limited to 50 outer updates, whereas CGS, FW, away-step FW, PG, and APG were limited to 1000 updates. Each figure reports the primal gap versus wall-clock time (seconds), the number of iterations, and the number of LMO calls. Because the optimal objective values for ridge-regularized logistic regression were unknown, we selected the best-performing methods, ran them for a sufficiently large number of iterations, and used the smallest objective value obtained.

4.1 Matrix Sensing with Squared Loss

Let $\mathbb{S}^n$ ($\mathbb{S}^n_+$) denote real symmetric (positive semidefinite) $n \times n$ matrices. Let $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ be the linear operator defined by $\mathcal{A}(X) = (\langle A_1, X\rangle, \dots, \langle A_m, X\rangle)^\mathsf{T}$, where $A_1, \dots, A_m \in \mathbb{R}^{n\times n}$ are given sensing matrices. We consider the matrix sensing problem with squared loss defined by $\min_{X\in\mathcal{M}_{R_\mathrm{Tr}}} \frac{1}{2}\|\mathcal{A}(X) - y_\mathrm{obs}\|^2$, where $\mathcal{M}_{R_\mathrm{Tr}} = \{X \in \mathbb{S}^n_+ \mid \operatorname{Tr} X = R_\mathrm{Tr}\}$ and $y_\mathrm{obs} \in \mathbb{R}^m$ is the vector of observations. We used a controlled synthetic construction that allowed us to independently prescribe the condition number of the quadratic objective and the alignment of the initial error with the high-curvature subspace. This enabled us to examine the sensitivity of first- and second-order methods to ill-conditioning. The details of the construction are given in Section E.1.

We set $n = 750$, $m = n(n+1)/2$, $R_\mathrm{Tr} = 1$, $\varphi = 16$, $\operatorname{rank} X_\star = 1$, and condition number $\Lambda = 1000$ and generated data using seed 2026. The initial point was $X_0 = (R_\mathrm{Tr}/n)I$. We omitted SOCGS because the convergence guarantees in Carderera & Pokutta (2026) do not cover $\mathcal{M}_{R_\mathrm{Tr}}$. Figure 1 shows the primal gap for $\varpi = 0.8$ (see Section E.1 for ϖ). Results for $\varpi \in \{0.2, 0.5\}$ appear in Section E.1. RBNFW (`local2`, `local3`) met their switching criteria by $k = 2$ and switched to full steps with the local inner criterion. RBNFW variants performed best in wall-clock time.

4.2 Ridge-Regularized Logistic Regression

Let $A \in \mathbb{R}^{m\times n}$ be a binary-classification design matrix, let a_i^T denote its ith row, and let $y_i \in \{-1, +1\}$ be the corresponding label. We solved the constrained ridge-regularized logistic regression problem $\min_{x\in\mathcal{X}} \frac{1}{m}\sum_{i=1}^m \log(1 + \exp(-y_i a_i^\mathsf{T} x)) + \frac{\beta}{2}\|x\|^2$. We consider two choices of the feasible set $\mathcal{X}$. The first is the ℓ_2-ball $\{x \in \mathbb{R}^n \mid \|x\| \le R\}$ with radius $R > 0$. The second is the sparse polytope $\{x \in \mathbb{R}^n \mid \|x\|_1 \le \xi R_\infty, \|x\|_\infty \le R_\infty\}$ with $R_\infty > 0$ and $\xi \in \{1, \dots, n\}$. We used Algorithm 1 for the ℓ_2-ball constraint and Algorithm 3 for the sparse polytope constraint.

We set $\beta = 10^{-3}$, $R = 1$, $\xi = 10$, and $R_\infty = 1/\sqrt{10}$, and initialized at $x_0 = 0$ for the ℓ_2-ball and at a vertex $x_0 \in \operatorname{argmin}_{x\in\mathcal{X}}\langle\nabla f(0), x\rangle$ for the sparse polytope. We used the a9a ($m = 32561$, $n = 123$), covtype.binary ($m = 581012$, $n = 54$), mushrooms ($m = 8124$, $n = 112$), phishing ($m = 11055$, $n = 68$), and w7a ($m = 24692$, $n = 300$)* datasets from the LIBSVM repository (Chang & Lin, 2011). Figures 2 and 3 show the primal gap for the covtype.binary dataset with ℓ_2-ball and sparse polytope constraints, respectively. Figure 4 shows the primal gap for the a9a dataset with a sparse polytope constraint. The RBNFW variants performed best in Figure 2. For the sparse polytope constraint, the wall-clock performance of the RBNFW variants was nearly identical to that of SOCGS in Figure 3. Results for the remaining datasets and constraints are reported in Section E.2. Across these datasets, RBNFW consistently exhibited stable and strong performance.

5 Conclusion and Future Work

Conclusion. We developed damped Newton FW methods for convex optimization over compact convex sets. Our methods approximately solve constrained damped Newton subproblems using only LMO calls, with either a standard FW or an away-step FW inner loop. We established global linear convergence with residual backtracking. Switching to full steps yields local superlinear or quadratic convergence. A projection-free Newton method can combine global linear convergence with local quadratic convergence over general compact convex sets, without requiring polyhedrality, strict complementarity, or a lower bound on the primal gap for setting the inner accuracy. In experiments on matrix sensing and ridge-regularized logistic regression, residual backtracking was consistently robust and often achieved the best wall-clock performance among the tested methods.

Future Work. Throughout this work, we form each Newton subproblem using the exact Hessian. An important direction for future work is to accommodate approximate Hessians or to reuse the same Hessian over multiple iterations through lazy updates, thereby reducing the computational cost of constructing and solving the subproblems. The proposed methods also require setting at least ω and ρ, and the theoretically admissible range of ω depends on problem constants and on the choice of ρ. A fully parameter-free method would therefore need to estimate the relevant problem constants adaptively and select the remaining algorithmic parameters automatically.

*The phishing and w7a datasets were used for sparse polytope constraints.

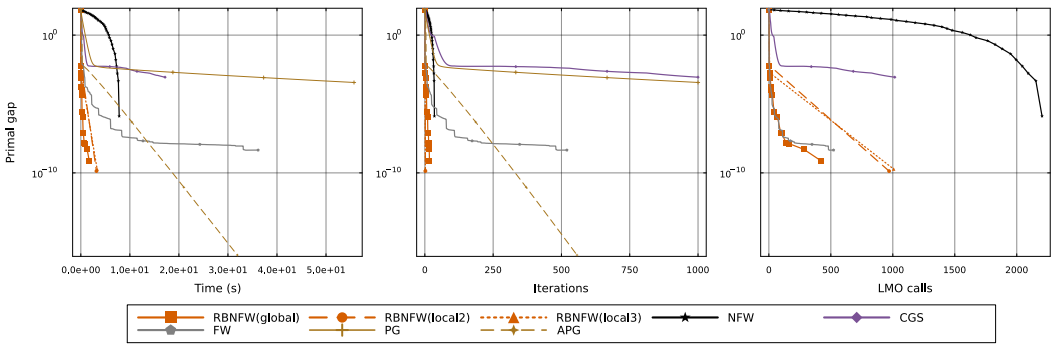


Figure 1: Primal gap for matrix sensing with $\varpi = 0.8$ using Algorithm 1.

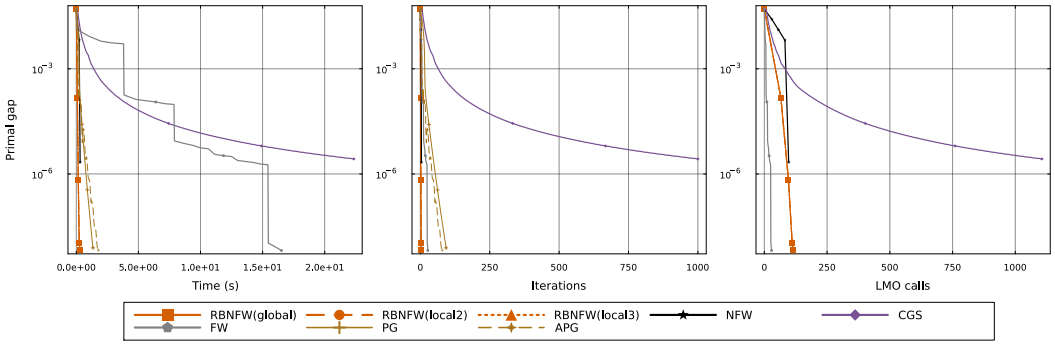


Figure 2: Primal gap for the covtype.binary dataset with an ℓ_2-ball constraint using Algorithm 1.

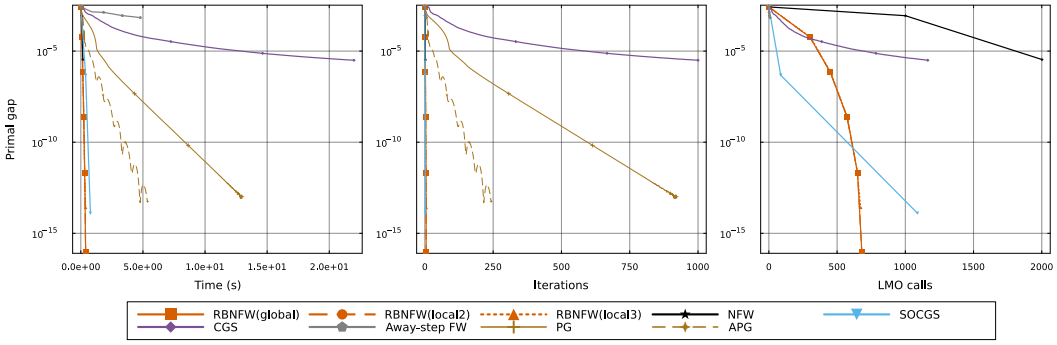


Figure 3: Primal gap for the covtype.binary dataset with a sparse polytope constraint using Algorithm 3.

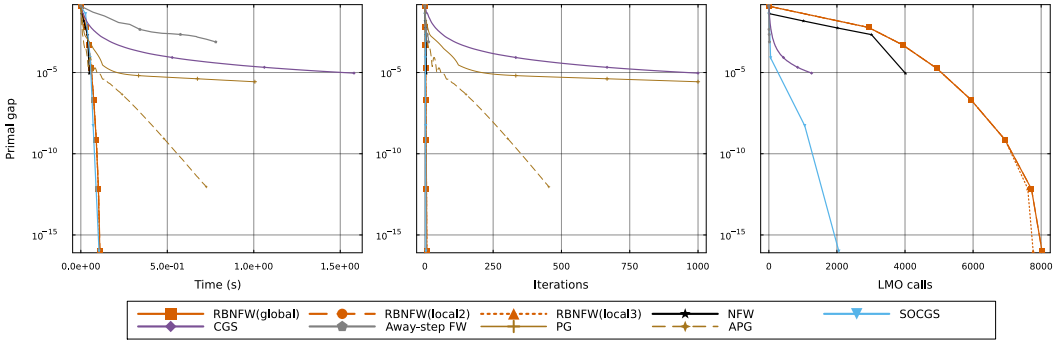


Figure 4: Primal gap for the a9a dataset with a sparse polytope constraint using Algorithm 3.

AI USE STATEMENT

In this work, we used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing key proof ingredients, assisting with proof writing, proposing or refining hypotheses, providing feedback on research methodology and experiments, implementing methods, assisting with translation, and interpreting results. We have not used generative AI tools for generating synthetic datasets or cleaning and reformatting datasets, and qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures, suggesting experimental parameters, writing or editing code, drafting manuscript text, summarizing and analyzing literature, searching for information, editing for readability, identifying relevant literature, and formatting references. We have reviewed all AI-assisted work. The code was checked by one author, and the proofs were verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work concerns the development and theoretical analysis of optimization algorithms, with numerical experiments on synthetic and existing benchmark datasets. We do not identify any specific ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

The algorithms and assumptions are described in Sections 2 and 3, with proofs provided in Sections B to D. The experimental settings and datasets are described in Sections 4 and E. The source code for the numerical experiments is included in the supplementary material.

REFERENCES

Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

Dimitri P Bertsekas. Projected Newton methods for optimization problems with simple constraints. *SIAM J. Control Optim.*, 20(2):221–246, 1982.

Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM J. Optim.*, 28(3):2131–2151, 2018.

Gábor Braun, Alejandro Carderera, Cyrille W Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. *Conditional gradient methods: From core principles to AI applications*, volume 35 of *MOS-SIAM Series on Optimization*. SIAM, Philadelphia, PA, 2025.

Ronald E Bruck. An iterative solution of a variational inequality for certain monotone operators in Hilbert space. *Bull. Am. Math. Soc.*, 81(5):890–892, 1975.

Michael D Canon and Clifton D Cullum. A tight upper bound on the rate of convergence of Frank–Wolfe algorithm. *SIAM J. Control*, 6(4):509–516, 1968.

Alejandro Carderera and Sebastian Pokutta. Second-order conditional gradient sliding. In Sanjeena Dang, Antoine Deza, Swati Gupta, Paul D McNicholas, Sebastian Pokutta, and Masashi Sugiyama (eds.), *Data science and optimization*, volume 91 of *Fields Institute Communications*, pp. 303–321. Springer, Cham, Switzerland, 2026.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011.

Cyrille W Combettes and Sebastian Pokutta. Complexity of linear minimization and projection on some sets. *Oper. Res. Lett.*, 49(4):565–571, 2021.

Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200, 2005.

Nikita Doikov and Yurii Nesterov. Contracting proximal methods for smooth convex optimization. *SIAM J. Optim.*, 30(4):3146–3169, 2020.

Nikita Doikov and Yurii Nesterov. Affine-invariant contracting-point methods for convex optimization. *Math. Program.*, 198(1):115–137, 2023.

Nikita Doikov and Yurii Nesterov. Gradient regularization of Newton method with Bregman distances. *Math. Program.*, 204(1–2):1–25, 2024.

Nikita Doikov, El Mahdi Chayti, and Martin Jaggi. Second-order optimization with lazy Hessians. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 8138–8161, Honolulu, Hawaii, USA, 2023. PMLR.

Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-universal regularized Newton method. *SIAM J. Optim.*, 34(1):27–56, 2024.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Nav. Res. Logist. Q.*, 3(1–2):95–110, 1956.

Kiwamu Fujiki, Shota Takahashi, and Akiko Takeda. Approximate Bregman proximal gradient algorithm with variable metric Armijo–Wolfe line search. *arXiv preprint arXiv:2510.06615*, 2025.

Dan Garber and Elad Hazan. Faster rates for the Frank–Wolfe method over strongly convex sets. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *PMLR*, pp. 541–549, Lille, France, 2015. PMLR.

Max L N Gonçalves and Jefferson G Melo. A Newton conditional gradient method for constrained nonlinear systems. *J. Comput. Appl. Math.*, 311:473–483, 2017.

Jacques Guélat and Patrice Marcotte. Some comments on Wolfe's 'away step'. *Math. Program.*, 35 (1):110–119, 1986.

Slavomír Hanzely, Dmitry Kamzolov, Dmitry Pasechnyuk, Alexander Gasnikov, Peter Richtárik, and Martin Takáč. A damped Newton method achieves global $\mathcal{O}(1/k^2)$ and local quadratic convergence rate. In Sanmi Koyejo, Shakir Mohamed, Alekh Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25320–25334, New Orleans, Louisiana, USA, 2022. Curran Associates, Inc.

Slavomír Hanzely, Farshed Abdukhakimov, and Martin Takáč. Newton method revisited: Global convergence rates up to $O(k^{-3})$ for stepsize schedules and linesearch procedures. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust (eds.), *The Fourteenth International Conference on Learning Representations*, Rio de Janeiro, Brazil, 2026.

Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *PMLR*, pp. 427–435, Atlanta, Georgia, USA, 2013. PMLR.

Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Projection-free optimization on uniformly convex sets. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pp. 19–27, Online, 2021. PMLR.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank–Wolfe optimization variants. In Corinna Cortes, Neil D Lawrence, Daniel D Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28, pp. 496–504, Montréal, Canada, 2015. Curran Associates, Inc.

Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM J. Optim.*, 26(2):1379–1409, 2016.

Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM J. Optim.*, 24(3):1420–1443, 2014.

Evgenii S Levitin and Boris T Polyak. Constrained minimization methods. *U.S.S.R. Comput. Math. Math. Phys.*, 6(5):1–50, 1966.

Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

Deyi Liu, Volkan Cevher, and Quoc Tran-Dinh. A Newton Frank–Wolfe method for constrained self-concordant minimization. *J. Global Optim.*, 83(2):273–299, 2022.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, 28(1):333–354, 2018.

Konstantin Mishchenko. Regularized Newton method with global $\mathcal{O}(1/k^2)$ convergence. *SIAM J. Optim.*, 33(3):1440–1462, 2023.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O\left(\frac{1}{k^2}\right)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

Yurii Nesterov. *Lectures on convex optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, Switzerland, 2018.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108(1):177–205, 2006.

Gergely Odor, Yen-Huan Li, Alp Yurtsever, Ya-Ping Hsieh, Quoc Tran-Dinh, Marwa El Halabi, and Volkan Cevher. Frank–Wolfe works for non-Lipschitz continuous gradient objectives: Scalable Poisson phase retrieval. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6230–6234, Shanghai, China, 2016. IEEE.

Gregory B Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *J. Math. Anal. Appl.*, 72(2):383–390, 1979.

Fabian Pedregosa, Geoffrey Négiar, Armin Askari, and Martin Jaggi. Linearly convergent Frank–Wolfe with backtracking line-search. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pp. 1–10, Online, 2020. PMLR.

Andrei Semenov, Martin Jaggi, and Nikita Doikov. Gradient-normalized smoothness for optimization with approximate Hessians. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust (eds.), *The Fourteenth International Conference on Learning Representations*, Rio de Janeiro, Brazil, 2026.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and Muons: Optimization via stochastic Frank–Wolfe under heavy-tailed noise. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *PMLR*, Seoul, South Korea, 2026. PMLR.

Shota Takahashi. Adaptive conditional gradient sliding: Projection-free and line-search-free acceleration. *arXiv preprint arXiv:2601.20443*, 2026.

Shota Takahashi and Akiko Takeda. Approximate Bregman proximal gradient algorithm for relatively smooth nonconvex optimization. *Comput. Optim. Appl.*, 90(1):227–256, 2025.

Shota Takahashi, Mituhiro Fukuda, and Mirai Tanaka. New Bregman proximal type algorithms for solving DC optimization problems. *Comput. Optim. Appl.*, 83(3):893–931, 2022.

Shota Takahashi, Mirai Tanaka, and Shiro Ikeda. Blind deconvolution with non-smooth regularization via Bregman proximal DCAs. *Signal Process.*, 202:108734, 2023.

Shota Takahashi, Sebastian Pokutta, and Akiko Takeda. Fast Frank–Wolfe algorithms with adaptive Bregman step-size for weakly convex functions. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust (eds.), *The Fourteenth International Conference on Learning Representations*, Rio de Janeiro, Brazil, 2026a.

Shota Takahashi, Mirai Tanaka, and Shiro Ikeda. Majorization-minimization Bregman proximal gradient algorithms for NMF with the Kullback–Leibler divergence. *J. Optim. Theory Appl.*, 208 (1):14, 2026b.

Haoyue Wang, Shibal Ibrahim, and Rahul Mazumder. Nonparametric finite mixture models with possible shape constraints: A cubic Newton approach. *SIAM J. Math. Data Sci.*, 7(1):163–188, 2025.

Philip Wolfe. Convergence theory in nonlinear programming. In Jean Abadie (ed.), *Integer and nonlinear programming*, pp. 1–36. North-Holland Publishing Company, Amsterdam, 1970.

Zev Woodstock. High-precision linear minimization is no slower than projection. *Optim. Lett.*, 20 (4):909–915, 2026.

A Additional Related Work

Projection and Proximal Methods. For smooth constrained optimization, projected gradient methods are widely used and can be viewed as proximal gradient methods in which the nonsmooth term is the indicator function of the feasible set. More generally, proximal gradient methods can be interpreted as forward–backward splitting schemes for minimizing the sum of a smooth function and a proximable nonsmooth function. The operator-splitting foundations of these methods trace back to early work on monotone operators and variational inequalities (Bruck, 1975; Lions & Mercier, 1979; Passty, 1979); see also (Combettes & Wajs, 2005) for a proximal forward–backward formulation in signal recovery. The fast iterative shrinkage-thresholding algorithm (FISTA) (Beck & Teboulle, 2009) is a widely used accelerated proximal gradient method. Under relative smoothness, the Euclidean distance can be replaced by the Bregman distance, and the Bregman proximal gradient method can be derived in this setting (Bauschke et al., 2017; Bolte et al., 2018; Lu et al., 2018). Bregman proximal methods have further been developed for difference-of-convex optimization (Takahashi et al., 2022) and signal-processing applications (Takahashi et al., 2023; 2026b). Approximate Bregman proximal gradient methods, which use second-order information from Bregman distances, have also been studied (Takahashi & Takeda, 2025; Fujiki et al., 2025).

Second-Order Methods. Classical Newton's method can attain local quadratic convergence under standard regularity conditions while its unit-step iteration need not be globally convergent. Common globalization mechanisms include damping, line searches, trust regions, and regularization. For convex objectives with Lipschitz-continuous Hessians, the cubic-regularized Newton method of Nesterov & Polyak (2006) achieves a global rate of $\mathcal{O}(k^{-2})$ and admits local quadratic convergence under standard local regularity conditions. More recent work has shown that similarly fast global guarantees can be obtained using damped or quadratically regularized Newton steps. In particular, Hanzely et al. (2022) proposed an explicit damping schedule that achieves a global $\mathcal{O}(k^{-2})$ rate together with local quadratic convergence, while Mishchenko (2023) established a global $\mathcal{O}(k^{-2})$ rate using gradient-dependent quadratic regularization. Gradient regularization has also been extended to non-Euclidean geometries using Bregman distances and to composite convex optimization (Doikov & Nesterov, 2024). The super-universal regularized Newton method (Doikov et al., 2024) adaptively handles several Hölder smoothness classes, and Hanzely et al. (2026) developed stepsize schedules and universal backtracking strategies for Newton methods under Hölder continuity of the Hessian or the third derivative. The latter work provides the main unconstrained motivation for the stepsize strategies developed in this paper. For composite optimization, proximal Newton methods minimize a quadratic model of the smooth term together with the nonsmooth term and preserve Newton-type local convergence with inexact subproblem solutions under suitable regularity and inner accuracy conditions (Lee et al., 2014). This framework includes convex constraints

by taking the nonsmooth term to be the indicator function of the feasible set. Our contribution is an LMO-based framework that combines computable FW-gap-based inner stopping criteria, residual backtracking, and a switch to full steps after finitely many iterations to establish global linear and fast local convergence over general compact convex sets. Contracting proximal methods instead minimize contracted objective models augmented by Bregman regularization and allow inexact solutions of the resulting auxiliary subproblems (Doikov & Nesterov, 2020). Closely related to our setting, the affine-invariant contracting-point framework of Doikov & Nesterov (2023) recovers the FW method in its first-order instance and yields an inexact Contracting Newton method in its second-order instance. This framework applies to bounded domains and admits an implementation based on LMOs, but its globalization mechanism contracts the domain, whereas our framework damps the quadratic model. In addition to NFW and SOCGS, Gonçalves & Melo (2017) proposed the Newton conditional gradient method, and Wang et al. (2025) proposed the cubic-regularized Newton method for self-concordant finite-mixture objectives with polyhedral constraints. Their cubic subproblems are solved using away-step FW methods, and their analysis establishes a global $\mathcal{O}(k^{-2})$ rate and local quadratic convergence. Another line of research reduces the cost of second-order information by reusing or approximating the Hessian. Doikov et al. (2023) proposed lazy Hessian updates that reuse the same Hessian over several iterations, while Semenov et al. (2026) developed gradient-regularized methods with approximate Hessians.

B Proofs of Section 2

B.1 Proof of Proposition 2.1

Proof of Proposition 2.1. If $D = 0$, then $\mathcal{X}$ is a singleton and no inner iteration is required. Hence, we assume that $D > 0$.

For any $u, v \in \mathcal{X}$ and $\gamma \in [0, 1]$, we have

$$\begin{aligned}\Psi_{\alpha_k}(u + \gamma(v - u); x_k) &= \Psi_{\alpha_k}(u; x_k) + \gamma \langle \nabla \Psi_{\alpha_k}(u; x_k), v - u \rangle + \frac{\gamma^2}{2\alpha_k} \|v - u\|_{x_k}^2 \\ &\leq \Psi_{\alpha_k}(u; x_k) + \gamma \langle \nabla \Psi_{\alpha_k}(u; x_k), v - u \rangle + \frac{2\gamma^2 D^2}{\alpha_k},\end{aligned}$$

where we used

$$\|v - u\|_{x_k} \leq \|v - x_k\|_{x_k} + \|u - x_k\|_{x_k} \leq 2D.$$

Since the stepsize γ_t in Algorithm 2 exactly minimizes $\Psi_{\alpha_k}(u; x_k)$ along the segment joining u_t and v_t, for any $\gamma \in [0, 1]$,

$$\Psi_{\alpha_k}(u_{t+1}; x_k) \leq \Psi_{\alpha_k}(u_t; x_k) - \gamma G_t + \frac{2\gamma^2 D^2}{\alpha_k}, \tag{B.1}$$

where G_t is the FW gap of $\Psi_{\alpha_k}(\cdot; x_k)$ over $\mathcal{X}$ at u_t (see (2.2)). Moreover, the convexity of $\Psi_{\alpha_k}(\cdot; x_k)$ gives

$$\Psi_{\alpha_k}(u_t; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \leq G_t. \tag{B.2}$$

Suppose that Algorithm 2 does not terminate at $t = 0$. Substituting the comparison value $\gamma = 1$ into (B.1) and using (B.2), we obtain

$$\Psi_{\alpha_k}(u_1; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \leq \frac{2D^2}{\alpha_k}. \tag{B.3}$$

We next show that, for every $t \geq 1$,

$$\Psi_{\alpha_k}(u_t; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \leq \frac{8D^2}{\alpha_k(t + 1)}. \tag{B.4}$$

The case $t = 1$ follows from (B.3). Suppose that (B.4) holds for some $t \geq 1$. Substituting the comparison value $\gamma = 2/(t + 2)$ into (B.1) and using (B.2), we obtain

$$\Psi_{\alpha_k}(u_{t+1}; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) = \min_{\gamma \in [0,1]} \Psi_{\alpha_k}((1 - \gamma)u_t + \gamma v_t; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k)$$

$$
\begin{aligned}
&\leq \Psi_{\alpha_k}\left(\left(1-\frac{2}{t+2}\right)u_t+\frac{2}{t+2}v_t;x_k\right)-\min_{u\in\mathcal{X}}\Psi_{\alpha_k}(u;x_k)\\
&=\Psi_{\alpha_k}(u_t;x_k)-\min_{u\in\mathcal{X}}\Psi_{\alpha_k}(u;x_k)-\frac{2}{t+2}G_t\\
&\quad+\frac{2}{\alpha_k(t+2)^2}\|v_t-u_t\|_{x_k}^2\\
&\overset{(B.2)}{\leq}\frac{t}{t+2}\left(\Psi_{\alpha_k}(u_t;x_k)-\min_{u\in\mathcal{X}}\Psi_{\alpha_k}(u;x_k)\right)\\
&\quad+\frac{2}{\alpha_k(t+2)^2}\|v_t-u_t\|_{x_k}^2\\
&\leq\frac{t}{t+2}\frac{8D^2}{\alpha_k(t+1)}+\frac{8D^2}{\alpha_k(t+2)^2}\leq\frac{8D^2}{\alpha_k(t+2)}.
\end{aligned}
$$

The last inequality follows from

$$
\frac{t}{(t+1)(t+2)}+\frac{1}{(t+2)^2}\leq\frac{1}{t+2}.
$$

Therefore, (B.4) follows by induction.

Let $U\geq 1$ be an integer. If Algorithm 2 terminates within the first $U>0$ inner updates, then the conclusion already holds. Hence, suppose that $u_1,\dots,u_{U+1}$ are generated. Setting $\gamma=2/(t+2)$ in (B.1), for every $t=1,\dots,U$, we obtain

$$
\frac{2}{t+2}G_t\leq\Psi_{\alpha_k}(u_t;x_k)-\Psi_{\alpha_k}(u_{t+1};x_k)+\frac{8D^2}{\alpha_k(t+2)^2}.
$$

Multiplying both sides by $(t+1)(t+2)/2$ and summing over $t=1,\dots,U$, we obtain

$$
\begin{aligned}
\sum_{t=1}^{U}(t+1)G_t&\leq\sum_{t=1}^{U}\frac{(t+1)(t+2)}{2}\left(\Psi_{\alpha_k}(u_t;x_k)-\Psi_{\alpha_k}(u_{t+1};x_k)\right)+\frac{4D^2}{\alpha_k}\sum_{t=1}^{U}\frac{t+1}{t+2}\\
&=3\left(\Psi_{\alpha_k}(u_1;x_k)-\min_{u\in\mathcal{X}}\Psi_{\alpha_k}(u;x_k)\right)\\
&\quad+\sum_{t=2}^{U}\left(\frac{(t+1)(t+2)}{2}-\frac{t(t+1)}{2}\right)\left(\Psi_{\alpha_k}(u_t;x_k)-\min_{u\in\mathcal{X}}\Psi_{\alpha_k}(u;x_k)\right)\\
&\quad-\frac{(U+1)(U+2)}{2}\left(\Psi_{\alpha_k}(u_{U+1};x_k)-\min_{u\in\mathcal{X}}\Psi_{\alpha_k}(u;x_k)\right)+\frac{4D^2}{\alpha_k}\sum_{t=1}^{U}\frac{t+1}{t+2}\\
&=3\left(\Psi_{\alpha_k}(u_1;x_k)-\min_{u\in\mathcal{X}}\Psi_{\alpha_k}(u;x_k)\right)\\
&\quad+\sum_{t=2}^{U}(t+1)\left(\Psi_{\alpha_k}(u_t;x_k)-\min_{u\in\mathcal{X}}\Psi_{\alpha_k}(u;x_k)\right)\\
&\quad-\frac{(U+1)(U+2)}{2}\left(\Psi_{\alpha_k}(u_{U+1};x_k)-\min_{u\in\mathcal{X}}\Psi_{\alpha_k}(u;x_k)\right)+\frac{4D^2}{\alpha_k}\sum_{t=1}^{U}\frac{t+1}{t+2}\\
&\overset{(a)}{\leq}3\frac{2D^2}{\alpha_k}+\sum_{t=2}^{U}(t+1)\frac{8D^2}{\alpha_k(t+1)}+\frac{4D^2}{\alpha_k}\sum_{t=1}^{U}\frac{t+1}{t+2}\\
&\leq\frac{6D^2}{\alpha_k}+\frac{8(U-1)D^2}{\alpha_k}+\frac{4UD^2}{\alpha_k}=\frac{(12U-2)D^2}{\alpha_k}\leq\frac{12UD^2}{\alpha_k},
\end{aligned}
$$

where (a) follows from (B.3) and (B.4). Consequently,

$$
\min_{1\leq t\leq U}G_t\frac{U(U+3)}{2}\leq\sum_{t=1}^{U}(t+1)G_t\leq\frac{12UD^2}{\alpha_k},
$$

and hence

$$\min_{1\le t\le U} G_t \le \frac{24D^2}{\alpha_k(U+3)}. \tag{B.5}$$

If Algorithm 2 does not terminate at $t=0$, set

$$U = \left\lceil \frac{24D^2}{\alpha_k \eta_k} \right\rceil.$$

Then, by (B.5), there exists some $t \in \{1,\dots,U\}$ such that $G_t \le \eta_k$. Therefore, Algorithm 2 terminates after at most U inner updates. □

B.2 Proof of Proposition 2.2

We present `AFW` of Algorithm 3 in Algorithm 5.

Algorithm 5: `AFW`$(h, H, u, \eta, \mathcal{S}, \lambda)$

$u_0 \leftarrow u$, $\mathcal{S}_0 \leftarrow \mathcal{S}$, $\lambda_0 \leftarrow \lambda$
for $t = 0, 1, 2, \dots$ **do**
  $v_t^{\mathrm{FW}} \leftarrow \operatorname{argmin}_{v\in\mathcal{X}} \langle h + H(u_t - u), v\rangle$
  if $\langle h + H(u_t - u), u_t - v_t^{\mathrm{FW}}\rangle \le \eta$ **then**
    return $(u_t, \mathcal{S}_t, \lambda_t)$
  $v_t^{\mathrm{A}} \leftarrow \operatorname{argmax}_{v\in\mathcal{S}_t} \langle h + H(u_t - u), v\rangle$
  if $\langle h + H(u_t - u), u_t - v_t^{\mathrm{FW}}\rangle \ge \langle h + H(u_t - u), v_t^{\mathrm{A}} - u_t\rangle$ **then**
    $d_t \leftarrow u_t - v_t^{\mathrm{FW}}$, $\gamma_{t,\max} \leftarrow 1$ ▷ `FW step`
  else
    $d_t \leftarrow v_t^{\mathrm{A}} - u_t$, $\gamma_{t,\max} \leftarrow \frac{\lambda_{v_t^{\mathrm{A}},t}}{1-\lambda_{v_t^{\mathrm{A}},t}}$ ▷ `Away step`
  $\gamma_t \leftarrow \min\left\{\gamma_{t,\max}, \frac{\langle h+H(u_t-u), d_t\rangle}{\|d_t\|_H^2}\right\}$
  $u_{t+1} \leftarrow u_t - \gamma_t d_t$
  if $\langle h + H(u_t - u), u_t - v_t^{\mathrm{FW}}\rangle \ge \langle h + H(u_t - u), v_t^{\mathrm{A}} - u_t\rangle$ **then**
    $\lambda_{v,t+1} \leftarrow (1-\gamma_t)\lambda_{v,t}$ for all $v \in \mathcal{S}_t \setminus \{v_t^{\mathrm{FW}}\}$ ▷ `FW step`
    $\lambda_{v_t^{\mathrm{FW}},t+1} \leftarrow \begin{cases} \gamma_t & \text{if } v_t^{\mathrm{FW}} \notin \mathcal{S}_t \\ (1-\gamma_t)\lambda_{v_t^{\mathrm{FW}},t} + \gamma_t & \text{if } v_t^{\mathrm{FW}} \in \mathcal{S}_t \end{cases}$
    $\mathcal{S}_{t+1} \leftarrow \begin{cases} \mathcal{S}_t \cup \{v_t^{\mathrm{FW}}\} & \text{if } \gamma_t < 1 \\ \{v_t^{\mathrm{FW}}\} & \text{if } \gamma_t = 1 \end{cases}$
  else
    $\lambda_{v,t+1} \leftarrow (1+\gamma_t)\lambda_{v,t}$ for $v \in \mathcal{S}_t \setminus \{v_t^{\mathrm{A}}\}$ ▷ `Away step`
    $\lambda_{v_t^{\mathrm{A}},t+1} \leftarrow (1+\gamma_t)\lambda_{v_t^{\mathrm{A}},t} - \gamma_t$
    $\mathcal{S}_{t+1} \leftarrow \begin{cases} \mathcal{S}_t & \text{if } \lambda_{v_t^{\mathrm{A}},t+1} > 0 \\ \mathcal{S}_t \setminus \{v_t^{\mathrm{A}}\} & \text{if } \lambda_{v_t^{\mathrm{A}},t+1} = 0 \end{cases}$

The weight vector $\lambda_t = (\lambda_{v,t})_{v\in\mathcal{S}_t}$ satisfies $u_t = \sum_{v\in\mathcal{S}_t} \lambda_{v,t} v$, $\lambda_{v,t} > 0$, and $\sum_{v\in\mathcal{S}_t} \lambda_{v,t} = 1$.

Proof of Proposition 2.2. By the direction-selection rule of Algorithm 5, we have

$$\langle \nabla\Psi_{\alpha_k}(u_t; x_k), d_t\rangle \ge \frac{1}{2}\langle \nabla\Psi_{\alpha_k}(u_t; x_k), v_t^{\mathrm{A}} - v_t^{\mathrm{FW}}\rangle.$$

The geometric strong convexity inequality associated with the pyramidal width (Braun et al., 2025, Lemma 2.27) provides

$$\Psi_{\alpha_k}(u_t; x_k) - \min_{u\in\mathcal{X}} \Psi_{\alpha_k}(u; x_k) \le \frac{\langle \nabla\Psi_{\alpha_k}(u_t; x_k), v_t^{\mathrm{A}} - v_t^{\mathrm{FW}}\rangle^2}{2\mu_k \delta_{\mathcal{X}}^2} \le \frac{2\langle \nabla\Psi_{\alpha_k}(u_t; x_k), d_t\rangle^2}{\mu_k \delta_{\mathcal{X}}^2}. \tag{B.6}$$

For any $\gamma \in [0, \gamma_{t,\max}]$, the quadratic structure of $\Psi_{\alpha_k}(\cdot; x_k)$ yields

$$\Psi_{\alpha_k}(u_t - \gamma d_t; x_k) = \Psi_{\alpha_k}(u_t; x_k) - \gamma \langle \nabla \Psi_{\alpha_k}(u_t; x_k), d_t \rangle + \frac{\gamma^2}{2\alpha_k} \|d_t\|_{x_k}^2. \tag{B.7}$$

Suppose first that the current update is an away step with $\gamma_t < \gamma_{t,\max}$ or an FW step with $\gamma_t < 1$. By Line 11 of Algorithm 5,

$$\Psi_{\alpha_k}(u_t; x_k) - \Psi_{\alpha_k}(u_{t+1}; x_k) = \frac{\alpha_k \langle \nabla \Psi_{\alpha_k}(u_t; x_k), d_t \rangle^2}{2\|d_t\|_{x_k}^2} \geq \frac{\langle \nabla \Psi_{\alpha_k}(u_t; x_k), d_t \rangle^2}{2L_k d_{\mathcal{X}}^2},$$

where we used $\|d_t\|_{x_k}^2 / \alpha_k \leq L_k \|d_t\|^2 \leq L_k d_{\mathcal{X}}^2$. Combining this inequality with (B.6), we obtain

$$\Psi_{\alpha_k}(u_t; x_k) - \Psi_{\alpha_k}(u_{t+1}; x_k) \geq \frac{\mu_k \delta_{\mathcal{X}}^2}{4L_k d_{\mathcal{X}}^2} \left(\Psi_{\alpha_k}(u_t; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \right).$$

Next, suppose that the current step is an FW step with $\gamma_t = 1$. By Line 11 of Algorithm 5,

$$\frac{\alpha_k \langle \nabla \Psi_{\alpha_k}(u_t; x_k), d_t \rangle}{\|d_t\|_{x_k}^2} \geq 1.$$

Hence, (B.7) implies

$$\begin{aligned} \Psi_{\alpha_k}(u_t; x_k) - \Psi_{\alpha_k}(u_{t+1}; x_k) &\geq \frac{1}{2} \langle \nabla \Psi_{\alpha_k}(u_t; x_k), d_t \rangle = \frac{G_t}{2} \\ &\geq \frac{1}{2} \left(\Psi_{\alpha_k}(u_t; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \right), \end{aligned}$$

where the last inequality follows from (B.2) due to the convexity of $\Psi_{\alpha_k}(\cdot; x_k)$. Since $\mu_k \leq L_k$ and $\delta_{\mathcal{X}} \leq d_{\mathcal{X}}$, we have $\mu_k \delta_{\mathcal{X}}^2 / (4L_k d_{\mathcal{X}}^2) \leq 1/4$. Therefore, in either case,

$$\Psi_{\alpha_k}(u_{t+1}; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \leq \left(1 - \frac{\mu_k \delta_{\mathcal{X}}^2}{4L_k d_{\mathcal{X}}^2} \right) \left(\Psi_{\alpha_k}(u_t; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \right). \tag{B.8}$$

Thus, (B.8) holds for every non-drop step. For a drop step, exact line minimization gives $\Psi_{\alpha_k}(u_{t+1}; x_k) \leq \Psi_{\alpha_k}(u_t; x_k)$.

Each drop step removes one vertex from the active set, whereas each FW step adds at most one vertex. Hence, among the first U inner updates, the number of drop steps is at most the number of FW steps plus $|\widehat{\mathcal{S}}_k| - 1$. Since every FW step is a non-drop step, the number of non-drop steps is at least $\max\{0, \lceil (U - |\widehat{\mathcal{S}}_k| + 1)/2 \rceil\}$. Consequently,

$$\begin{aligned} \Psi_{\alpha_k}(u_U; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \leq &\left(\Psi_{\alpha_k}(u_0; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \right) \\ &\times \left(1 - \frac{\mu_k \delta_{\mathcal{X}}^2}{4L_k d_{\mathcal{X}}^2} \right)^{\max\{0, \lceil (U - |\widehat{\mathcal{S}}_k| + 1)/2 \rceil\}}. \end{aligned} \tag{B.9}$$

For every $\gamma \in [0, 1]$, we have

$$\begin{aligned} \Psi_{\alpha_k}(u_t; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) &\geq \Psi_{\alpha_k}(u_t; x_k) - \Psi_{\alpha_k}\big(u_t + \gamma(v_t^{\mathrm{FW}} - u_t); x_k\big) \\ &\geq \gamma G_t - \frac{L_k \gamma^2 d_{\mathcal{X}}^2}{2}. \end{aligned}$$

Suppose that $\Psi_{\alpha_k}(u_t; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \leq L_k d_{\mathcal{X}}^2 / 2$. If $G_t > L_k d_{\mathcal{X}}^2$, then choosing $\gamma = 1$ in the preceding inequality gives a contradiction. Hence, $G_t \leq L_k d_{\mathcal{X}}^2$, and choosing $\gamma = G_t / (L_k d_{\mathcal{X}}^2)$ yields

$$G_t \leq d_{\mathcal{X}} \sqrt{2L_k \left(\Psi_{\alpha_k}(u_t; x_k) - \min_{u \in \mathcal{X}} \Psi_{\alpha_k}(u; x_k) \right)}.$$

Therefore, $\Psi_{\alpha_k}(u_t; x_k) - \min_{u\in\mathcal{X}} \Psi_{\alpha_k}(u; x_k) \le \zeta_k$ implies $G_t \le d_{\mathcal{X}}\sqrt{2L_k\zeta_k} \le \eta_k$.

Finally, set

$$U := |\widehat{\mathcal{S}}_k| - 1 + 2\left\lceil \frac{\log\max\{1, g_k/\zeta_k\}}{-\log\left(1-\mu_k\delta_{\mathcal{X}}^2/(4L_k d_{\mathcal{X}}^2)\right)} \right\rceil.$$

Since $u_0 = x_k$ and the quadratic term in $\Psi_{\alpha_k}(\cdot; x_k)$ is nonnegative, we have the initial model primal-gap bound as follows:

$$\Psi_{\alpha_k}(x_k; x_k) - \min_{u\in\mathcal{X}} \Psi_{\alpha_k}(u; x_k) \le \max_{u\in\mathcal{X}} \langle \nabla f(x_k), x_k - u\rangle = g_k.$$

If the inner loop has not terminated earlier, then (B.9) and the initial model primal-gap bound give

$$\Psi_{\alpha_k}(u_U; x_k) - \min_{u\in\mathcal{X}} \Psi_{\alpha_k}(u; x_k) \le g_k\left(1 - \frac{\mu_k\delta_{\mathcal{X}}^2}{4L_k d_{\mathcal{X}}^2}\right)^{\left\lceil \frac{\log\max\{1, g_k/\zeta_k\}}{-\log\left(1-\mu_k\delta_{\mathcal{X}}^2/(4L_k d_{\mathcal{X}}^2)\right)} \right\rceil} \le \zeta_k.$$

Hence, $G_U \le \eta_k$, and the inner loop terminates after at most U iterations. □

C Proofs of Section 3.1

Convergence Terminology. For a nonnegative sequence $b_k \to 0$, we say that it converges Q-linearly if $b_{k+1} \le \psi b_k$ eventually for some $\psi \in (0,1)$, and Q-converges with order at least $l > 1$ if $b_{k+1} \le C b_k^l$ eventually for some $C > 0$. We say that b_k converges R-linearly if $b_k \le C\psi^k$ for some $C > 0$ and $\psi \in (0,1)$. Q-quadratic convergence means Q-order at least two. For convergence of the iterates and the primal gap, these definitions apply to the scalar sequences $\|x_k - x^\star\|$ and $f(x_k) - f(x^\star)$, respectively.

C.1 Proof of Lemma 3.2

From the definition of w_k, we have

$$x_{k+1} - x_k = -\alpha_k \nabla^2 f(x_k)^{-1} w_k, \qquad \|x_{k+1} - x_k\|_{x_k} = \alpha_k \|w_k\|_{x_k}^*. \tag{C.1}$$

Proof of Lemma 3.2. For $k \ge 1$, applying (2.2) at iteration k with $v = x_k$ and at iteration $k-1$ with $v = x_{k+1}$ gives

$$\langle s_k, x_k - x_{k+1}\rangle \le \eta_k, \qquad \langle s_{k-1}, x_{k+1} - x_k\rangle \le \eta_{k-1}.$$

The same argument applies when $k = 0$ by setting $s_{-1} = 0$ and $\eta_{-1} = 0$. Therefore,

$$\langle s_k - s_{k-1}, x_{k+1} - x_k\rangle \ge -(\eta_k + \eta_{k-1}).$$

Using $s_k - s_{k-1} = w_k - r_k$ and (C.1), we obtain

$$\begin{aligned}\|w_k\|_{x_k}^{*2} &\le \left\langle r_k, \nabla^2 f(x_k)^{-1} w_k\right\rangle + \frac{\eta_k + \eta_{k-1}}{\alpha_k} \\ &\le \|r_k\|_{x_k}^* \|w_k\|_{x_k}^* + \frac{\eta_k + \eta_{k-1}}{\alpha_k},\end{aligned}$$

where the second inequality follows from the Cauchy–Schwarz inequality. This proves (3.4).

Suppose now that (3.2) holds at every iteration and that $\alpha_k \ge (1 + B_q \Delta_k^{\frac{q-2}{q-1}})^{-1}$. For $k \ge 1$ with $\Delta_{k-1} > 0$, set $t = \Delta_k/\Delta_{k-1} \ge \rho$. Then

$$\frac{\eta_{k-1}}{\eta_k} = \frac{\Delta_{k-1}^2(1 + B_q\Delta_k^{\frac{q-2}{q-1}})}{\Delta_k^2(1 + B_q\Delta_{k-1}^{\frac{q-2}{q-1}})} = t^{-2}\frac{1 + B_q t^{\frac{q-2}{q-1}}\Delta_{k-1}^{\frac{q-2}{q-1}}}{1 + B_q\Delta_{k-1}^{\frac{q-2}{q-1}}} \le \rho^{-2}. \tag{C.2}$$

Indeed, the last fraction is at most 1 when $t \in [\rho, 1]$ and at most $t^{\frac{q-2}{q-1}}$ when $t \ge 1$. In the latter case, the entire expression is at most $t^{-\frac{q}{q-1}} \le 1$. Thus, $\eta_{k-1} \le \rho^{-2}\eta_k$; the same inequality holds for $k = 0$ or $\Delta_{k-1} = 0$, since then $\eta_{k-1} = 0$. Consequently,

$$\frac{\eta_k + \eta_{k-1}}{\alpha_k} \le (1 + \rho^{-2})\left(1 + B_q\Delta_k^{\frac{q-2}{q-1}}\right)\eta_k = \omega(1 + \rho^{-2})\Delta_k^2.$$

Combining this with (3.4) and $\|r_k\|_{x_k}^* \le \Delta_k$ gives

$$\|w_k\|_{x_k}^{*2} \le \Delta_k \|w_k\|_{x_k}^* + \omega(1+\rho^{-2})\Delta_k^2.$$

Solving this quadratic inequality yields $\|w_k\|_{x_k}^* \le \kappa_\omega \Delta_k$. The same argument applies to each trial after replacing $(\alpha_k, x_{k+1}, s_k, w_k)$ by $(\alpha_{k,j}, x_{k,j}, s_{k,j}, w_{k,j})$, while keeping the previous accepted s_{k-1} and η_{k-1} fixed. In particular,

$$\|w_{k,j}\|_{x_k}^{*2} \le \|r_k\|_{x_k}^* \|w_{k,j}\|_{x_k}^* + \frac{\eta_k + \eta_{k-1}}{\alpha_{k,j}} \le \Delta_k \|w_{k,j}\|_{x_k}^* + \omega(1+\rho^{-2})\Delta_k^2, \tag{C.3}$$

because every RBNFW trial satisfies $\alpha_{k,j} \ge (1 + B_q \Delta_k^{\frac{q-2}{q-1}})^{-1}$. □

C.2 A Sufficient Condition for Acceptance

We now present a sufficient condition for the acceptance of a trial step in RBNFW.

Proposition C.1. Let $\|w_k\|_{x_k}^* > 0$, $p \in \{2,3\}$, $\nu \in [0,1]$, $q = p + \nu$, and $\alpha_k = 1/(1+\theta_k)$ with $\theta_k > 0$. Suppose that Assumption 3.1 holds and that $x_{k+1} \in \mathcal{X}$ satisfies (2.2). If

$$\theta_k \ge (9L_{p,\nu})^{\frac{1}{q-1}} \|w_k\|_{x_k}^{*\frac{q-2}{q-1}}, \tag{C.4}$$

then

$$\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \rangle \ge \frac{c_p}{\alpha_k \theta_k} \|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2}. \tag{C.5}$$

Proof. By Lemma C.2, (C.4) implies the assumptions of Lemma C.4. Applying that lemma and using $1 - \alpha_k = \alpha_k \theta_k$ yields (C.5). □

We first establish three auxiliary lemmas.

Lemma C.2. Let $p \in \{2,3\}$, $\nu \in [0,1]$, and $q = p + \nu$, and let $\alpha_k = 1/(1+\theta_k)$ with $\theta_k > 0$. Suppose also that $\|w_k\|_{x_k}^* > 0$ and that (C.4) holds. Then, the following statements hold.

(i) If $p = 2$, then

$$\theta_k \ge \frac{L_{2,\nu}}{1+\nu} \alpha_k^\nu \|w_k\|_{x_k}^{*\nu}.$$

(ii) If $p = 3$, then

$$\theta_k \ge \alpha_k \|w_k\|_{x_k}^* \max\left\{ 6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}, \frac{\sqrt{3} L_{3,\nu}}{(1+\nu)(2+\nu)} \alpha_k^\nu \|w_k\|_{x_k}^{*\nu} \right\}.$$

Proof. (i) The case $p = 2$: Since $q = 2 + \nu$, (C.4) gives

$$\theta_k^{1+\nu} \ge 9 L_{2,\nu} \|w_k\|_{x_k}^{*\nu}.$$

Therefore, using $\alpha_k \theta_k \le 1$, we obtain

$$\frac{L_{2,\nu}}{1+\nu} \alpha_k^\nu \|w_k\|_{x_k}^{*\nu} \le \frac{\alpha_k^\nu \theta_k^{1+\nu}}{9(1+\nu)} = \frac{(\alpha_k \theta_k)^\nu}{9(1+\nu)} \theta_k \le \theta_k.$$

(ii) The case $p = 3$: Since $q = 3 + \nu$, we have

$$\theta_k^{2+\nu} \ge 9 L_{3,\nu} \|w_k\|_{x_k}^{*(1+\nu)}. \tag{C.6}$$

First,

$$\left(6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k \|w_k\|_{x_k}^*\right)^{1+\nu} \le \theta_k^{1+\nu} \frac{6^{1+\nu}}{9(1+\nu)} \frac{\theta_k}{(1+\theta_k)^{1+\nu}} \le \theta_k^{1+\nu} \frac{6^{1+\nu}}{9(1+\nu)^2} \le \theta_k^{1+\nu},$$

where we used

$$\frac{\theta_k}{(1+\theta_k)^{1+\nu}} \le \frac{1}{1+\nu}, \qquad 6^{1+\nu} \le 9(1+\nu)^2.$$

Indeed, Bernoulli's inequality gives $(1+\theta_k)^{1+\nu} \ge 1+(1+\nu)\theta_k \ge (1+\nu)\theta_k > 0$. Taking reciprocals and multiplying by θ_k yields

$$\frac{\theta_k}{(1+\theta_k)^{1+\nu}} \le \frac{\theta_k}{(1+\nu)\theta_k} = \frac{1}{1+\nu}.$$

Moreover, $6^{1+\nu} \le 9(1+\nu)^2$ holds for every $\nu \in [0,1]$. Hence,

$$\theta_k \ge 6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k \|w_k\|_{x_k}^*.$$

In addition, (C.6) and $\alpha_k\theta_k \le 1$ imply

$$\frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(2+\nu)}\alpha_k^{1+\nu}\|w_k\|_{x_k}^{*(1+\nu)} \le \theta_k \frac{\sqrt{3}}{9(1+\nu)(2+\nu)}(\alpha_k\theta_k)^{1+\nu} \le \theta_k.$$

This proves the claim. □

Lemma C.3. Suppose that Assumption 3.1 holds. If $p=2$, then, for any $x, y \in \mathcal{X}$,

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\|_x^* \le \frac{L_{2,\nu}}{1+\nu}\|y-x\|_x^{1+\nu}.$$

If $p=3$, then, for any $x, y \in \mathcal{X}$,

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x) - \frac{1}{2}\nabla^3 f(x)[y-x, y-x]\|_x^* \le \frac{L_{3,\nu}}{(1+\nu)(2+\nu)}\|y-x\|_x^{2+\nu}.$$

Proof. Let $d \coloneqq y - x$. Since $\mathcal{X}$ is convex, $x + td \in \mathcal{X}$ for every $t \in [0,1]$. Suppose first that $\nu > 0$. If $p=2$, the fundamental theorem of calculus gives

$$\nabla f(y) - \nabla f(x) - \nabla^2 f(x)d = \int_0^1 \left(\nabla^2 f(x+td) - \nabla^2 f(x)\right) d\, dt.$$

Hence, by Assumption 3.1,

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)d\|_x^* \le L_{2,\nu}\|d\|_x^{1+\nu}\int_0^1 t^\nu\, dt = \frac{L_{2,\nu}}{1+\nu}\|d\|_x^{1+\nu}.$$

If $p=3$, the integral Taylor formula gives

$$\nabla f(y) - \nabla f(x) - \nabla^2 f(x)d - \frac{1}{2}\nabla^3 f(x)[d,d] = \int_0^1 (1-t)\left(\nabla^3 f(x+td) - \nabla^3 f(x)\right)[d,d]\, dt.$$

Therefore,

$$\begin{aligned}&\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)d - \tfrac{1}{2}\nabla^3 f(x)[d,d]\|_x^* \\ &\quad\le L_{3,\nu}\|d\|_x^{2+\nu}\int_0^1 (1-t)t^\nu dt = \frac{L_{3,\nu}}{(1+\nu)(2+\nu)}\|d\|_x^{2+\nu}.\end{aligned}$$

When $\nu = 0$, the same bounds follow by applying the one-dimensional mean-value theorem for $p=2$ and the second-order Taylor theorem for $p=3$ to $t \mapsto \langle \nabla f(x+td), u\rangle$ for arbitrary u with $\|u\|_x \le 1$, and then taking the supremum over such u. □

We next establish the counterpart of (Hanzely et al., 2026, Lemma 4).

Lemma C.4. Let $\|w_k\|_{x_k}^* > 0$, and let $\alpha_k = 1/(1+\theta_k)$ with $\theta_k > 0$.

(i) Suppose that Assumption 3.1 holds with $p = 2$ and that

$$\theta_k \geq \frac{L_{2,\nu}}{1+\nu}\alpha_k^\nu \|w_k\|_{x_k}^{*\nu}.$$

Then,

$$\left\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \right\rangle \geq \frac{1}{2(1-\alpha_k)} \|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2}.$$

(ii) Suppose that Assumption 3.1 holds with $p = 3$ and that

$$\theta_k \geq \alpha_k \|w_k\|_{x_k}^* \max\left\{ 6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}, \frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(2+\nu)}\alpha_k^\nu \|w_k\|_{x_k}^{*\nu} \right\}. \tag{C.7}$$

Then,

$$\left\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \right\rangle \geq \frac{1}{4(1-\alpha_k)} \|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2}.$$

Proof. (i) The case $p = 2$: By the $L_{2,\nu}$-Hölder continuity, Lemma C.3 implies

$$\begin{aligned}
\left(\frac{L_{2,\nu}}{1+\nu}\|x_{k+1} - x_k\|_{x_k}^{1+\nu}\right)^2 &\geq \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k)\|_{x_k}^{*2} \\
&= \|\nabla f(x_{k+1}) - \nabla f(x_k) + \alpha_k w_k\|_{x_k}^{*2} \\
&= \|\nabla f(x_{k+1}) + s_k - (1-\alpha_k)(\nabla f(x_k) + s_k)\|_{x_k}^{*2} \\
&= \|\nabla f(x_{k+1}) + s_k - (1-\alpha_k) w_k\|_{x_k}^{*2} \\
&= \|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2} + (1-\alpha_k)^2 \|w_k\|_{x_k}^{*2} \\
&\quad - 2(1-\alpha_k)\left\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \right\rangle.
\end{aligned}$$

It is therefore sufficient to show that

$$\frac{1-\alpha_k}{2}\|w_k\|_{x_k}^{*2} \geq \frac{1}{2(1-\alpha_k)}\left(\frac{L_{2,\nu}}{1+\nu}\|x_{k+1} - x_k\|_{x_k}^{1+\nu}\right)^2. \tag{C.8}$$

By (C.1) and $1 - \alpha_k = \alpha_k \theta_k$, (C.8) is equivalent to $\theta_k \geq \frac{L_{2,\nu}}{1+\nu}\alpha_k^\nu \|w_k\|_{x_k}^{*\nu}$, which is equivalent to the assumption of the lemma. Consequently,

$$\begin{aligned}
\left\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \right\rangle &\geq \frac{1}{2(1-\alpha_k)}\|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2} + \frac{1-\alpha_k}{2}\|w_k\|_{x_k}^{*2} \\
&\quad - \frac{1}{2(1-\alpha_k)}\left(\frac{L_{2,\nu}}{1+\nu}\|x_{k+1} - x_k\|_{x_k}^{1+\nu}\right)^2 \\
&\overset{\text{(C.8)}}{\geq} \frac{1}{2(1-\alpha_k)}\|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2}.
\end{aligned}$$

(ii) The case $p = 3$: First, by the $L_{3,\nu}$-Hölder continuity, Lemma C.3 implies

$$\begin{aligned}
&\left(\frac{L_{3,\nu}}{(1+\nu)(2+\nu)}\|x_{k+1} - x_k\|_{x_k}^{2+\nu}\right)^2 \\
&\geq \left\| \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k) - \frac{1}{2}\nabla^3 f(x_k)[x_{k+1} - x_k, x_{k+1} - x_k] \right\|_{x_k}^{*2} \\
&= \left\| \nabla f(x_{k+1}) + s_k - (1-\alpha_k) w_k - \frac{1}{2}\nabla^3 f(x_k)[x_{k+1} - x_k, x_{k+1} - x_k] \right\|_{x_k}^{*2} \\
&= \|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2} + (1-\alpha_k)^2 \|w_k\|_{x_k}^{*2} + \frac{1}{4}\|\nabla^3 f(x_k)[x_{k+1} - x_k, x_{k+1} - x_k]\|_{x_k}^{*2} \\
&\quad - 2(1-\alpha_k)\left\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \right\rangle
\end{aligned}$$

$$
\begin{aligned}
&+ (1-\alpha_k)\left\langle \nabla^2 f(x_k)^{-1/2} w_k, \nabla^2 f(x_k)^{-1/2} \nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k] \right\rangle \\
&- \left\langle \nabla^2 f(x_k)^{-1/2} (\nabla f(x_{k+1}) + s_k), \nabla^2 f(x_k)^{-1/2} \nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k] \right\rangle \\
&\overset{\text{(a)}}{\geq} \frac{1}{2}\|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2} + (1-\alpha_k)^2 \|w_k\|_{x_k}^{*2} - \frac{1}{4}\|\nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k]\|_{x_k}^{*2} \\
&- 2(1-\alpha_k)\left\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \right\rangle \\
&- (1-\alpha_k)\|w_k\|_{x_k}^{*} \|\nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k]\|_{x_k}^{*} \\
&\overset{\text{(b)}}{\geq} \frac{1}{2}\|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2} + (1-\alpha_k)^2 \|w_k\|_{x_k}^{*2} - \left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{2}{1+\nu}} \alpha_k^4 \|w_k\|_{x_k}^{*4} \\
&- 2(1-\alpha_k)\left\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \right\rangle \\
&- 2\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k^2 (1-\alpha_k)\|w_k\|_{x_k}^{*3}.
\end{aligned}
$$

In (a), we used the Cauchy–Schwarz and Young inequalities:

$$
\begin{aligned}
&- \left\langle \nabla^2 f(x_k)^{-1/2} (\nabla f(x_{k+1}) + s_k), \nabla^2 f(x_k)^{-1/2} \nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k] \right\rangle \\
&\geq -\|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*} \|\nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k]\|_{x_k}^{*} \\
&\geq -\frac{1}{2}\|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2} - \frac{1}{2}\|\nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k]\|_{x_k}^{*2},
\end{aligned}
$$

together with

$$
\begin{aligned}
&(1-\alpha_k)\left\langle \nabla^2 f(x_k)^{-1/2} w_k, \nabla^2 f(x_k)^{-1/2} \nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k] \right\rangle \\
&\geq -(1-\alpha_k)\|w_k\|_{x_k}^{*} \|\nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k]\|_{x_k}^{*}.
\end{aligned}
$$

In step (b), we used the definition of $\Gamma_{3,\mathcal{X}}$, the normalization of $L_{3,\nu}$ in Assumption 3.1, and (C.1):

$$
\begin{aligned}
\|\nabla^3 f(x_k)[x_{k+1}-x_k, x_{k+1}-x_k]\|_{x_k}^{*} &\leq \Gamma_{3,\mathcal{X}} \|x_{k+1}-x_k\|_{x_k}^2 \\
&\leq 2\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \|x_{k+1}-x_k\|_{x_k}^2 \\
&= 2\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k^2 \|w_k\|_{x_k}^{*2}.
\end{aligned}
$$

Rearranging the above inequality and using (C.1) gives

$$
\begin{aligned}
&\left\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \right\rangle \\
&\geq \frac{1}{4(1-\alpha_k)}\|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2} + \frac{1-\alpha_k}{2}\|w_k\|_{x_k}^{*2} - \frac{1}{2}\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{2}{1+\nu}} \frac{\alpha_k^4}{1-\alpha_k}\|w_k\|_{x_k}^{*4} \\
&\quad - \left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k^2 \|w_k\|_{x_k}^{*3} - \frac{1}{2(1-\alpha_k)}\left(\frac{L_{3,\nu}}{(1+\nu)(2+\nu)}\right)^2 (\alpha_k \|w_k\|_{x_k}^{*})^{2(2+\nu)}.
\end{aligned}
$$

If $\alpha_k = 1/(1+\theta_k)$ satisfies

$$
\frac{1-\alpha_k}{6}\|w_k\|_{x_k}^{*2} \geq \frac{1}{2}\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{2}{1+\nu}} \frac{\alpha_k^4}{1-\alpha_k}\|w_k\|_{x_k}^{*4}, \tag{C.9}
$$

$$
\frac{1-\alpha_k}{6}\|w_k\|_{x_k}^{*2} \geq \left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k^2 \|w_k\|_{x_k}^{*3}, \tag{C.10}
$$

$$
\frac{1-\alpha_k}{6}\|w_k\|_{x_k}^{*2} \geq \frac{1}{2(1-\alpha_k)}\left(\frac{L_{3,\nu}}{(1+\nu)(2+\nu)}\right)^2 (\alpha_k \|w_k\|_{x_k}^{*})^{2(2+\nu)}, \tag{C.11}
$$

then

$$
\left\langle \nabla f(x_{k+1}) + s_k, \nabla^2 f(x_k)^{-1} w_k \right\rangle \geq \frac{1}{4(1-\alpha_k)}\|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2}.
$$

Since $1-\alpha_k=\alpha_k\theta_k$, we have

$$1-\alpha_k=\alpha_k\theta_k \geq \begin{cases} \sqrt{3}\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\alpha_k^2\|w_k\|_{x_k}^*, & \text{if } \theta_k \geq \sqrt{3}\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\alpha_k\|w_k\|_{x_k}^*, \\ 6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\alpha_k^2\|w_k\|_{x_k}^*, & \text{if } \theta_k \geq 6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\alpha_k\|w_k\|_{x_k}^*, \\ \frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(2+\nu)}\alpha_k^{2+\nu}\|w_k\|_{x_k}^{*(1+\nu)}, & \text{if } \theta_k \geq \frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(2+\nu)}\alpha_k^{1+\nu}\|w_k\|_{x_k}^{*(1+\nu)}. \end{cases}$$

Hence, (C.7) implies (C.9), (C.10), and (C.11). □

C.3 Proof of Theorem 3.3

Proof of Theorem 3.3. If $w_k=0$, then (C.1) gives $x_{k+1}=x_k$, and hence $r_{k+1}=\nabla f(x_{k+1})+s_k=w_k=0$. Thus, (3.5) holds trivially. We therefore assume that $w_k\neq 0$.

For the root Newton rule and the accuracy, Lemma 3.2 gives

$$\theta_k=B_q\Delta_k^{\frac{q-2}{q-1}}=(9L_{p,\nu})^{\frac{1}{q-1}}(\kappa_\omega\Delta_k)^{\frac{q-2}{q-1}}\geq(9L_{p,\nu})^{\frac{1}{q-1}}\|w_k\|_{x_k}^{*\frac{q-2}{q-1}}, \tag{C.12}$$

so Proposition C.1 gives (C.5). It is not necessary that earlier updates use the root stepsize. For an update accepted by Algorithm 4, (C.5) follows directly from the acceptance condition (3.3). In both cases, $0<\theta_k\leq B_q\Delta_k^{\frac{q-2}{q-1}}$. Using the convexity of f, $\nabla f(x_{k+1})=r_{k+1}-s_k$, and $x_{k+1}-x_k=-\alpha_k\nabla^2 f(x_k)^{-1}w_k$, we obtain

$$\begin{aligned} f(x_k)-f(x_{k+1}) &\geq -\langle\nabla f(x_{k+1}),x_{k+1}-x_k\rangle \\ &= \alpha_k\langle r_{k+1},\nabla^2 f(x_k)^{-1}w_k\rangle-\langle s_k,x_k-x_{k+1}\rangle \\ &\overset{(2.2)}{\geq} \alpha_k\langle r_{k+1},\nabla^2 f(x_k)^{-1}w_k\rangle-\eta_k \\ &\overset{(C.5)}{\geq} \frac{c_p}{\theta_k}\|r_{k+1}\|_{x_k}^{*2}-\eta_k \\ &\geq \frac{c_p}{B_q}\frac{\|r_{k+1}\|_{x_k}^{*2}}{\Delta_k^{\frac{q-2}{q-1}}}-\eta_k. \end{aligned}$$

This proves (3.5) for both choices. □

C.4 Proof of Corollary 3.4

Proof of Corollary 3.4. Fix a nonterminal iteration k, so $\Delta_k>0$ and $\bar\theta_k>0$. Since $\nabla^2 f(x_k)$ is a fixed positive definite matrix and $\mathcal{X}$ is compact, $D_k:=\max_{u,v\in\mathcal{X}}\|u-v\|_{x_k}<\infty$. The proof of Proposition 2.1 applies with D replaced by D_k, so $\eta_k>0$ and $\alpha_{k,j}>0$ guarantee finite termination of each FW inner solve. For the away-step FW inner loop, fix any $\alpha\in(0,1]$ and $\eta>0$, and let G_t be the FW gap of $\Psi_\alpha(\cdot;x_k)$. Since k is nonterminal, $D_k>0$. By the direction-selection rule of Algorithm 5, we have $\langle\nabla\Psi_\alpha(u_t;x_k),d_t\rangle\geq G_t$ and $\|d_t\|_{x_k}\leq D_k$. The quadratic expansion (B.7) with $\alpha_k=\alpha$ and exact line minimization give

$$\Psi_\alpha(u_t;x_k)-\Psi_\alpha(u_{t+1};x_k)\geq\begin{cases}\dfrac{\alpha G_t^2}{2D_k^2}, & \gamma_t<\gamma_{t,\max},\\ \dfrac{G_t}{2}, & \text{FW step with } \gamma_t=1.\end{cases}$$

Drop steps do not increase the model. Each drop step removes one point from the active set, whereas each FW step adds at most one point. Hence, the number of drop steps is at most the number of FW steps plus $|\mathcal{S}_0|-1$. Since $\mathcal{S}_0$ is finite, an infinite inner sequence would have infinitely many non-drop steps. If $G_t>\eta$ throughout, each such step would decrease the model by at least $\min\{\alpha\eta^2/(2D_k^2),\eta/2\}>0$, contradicting its lower boundedness on $\mathcal{X}$. Thus, the AFW inner loop terminates for every $\alpha\in(0,1]$ and $\eta>0$, in particular for $\alpha=\alpha_{k,j}$ and $\eta=\eta_k$.

If no earlier trial is accepted, the search reaches a finite index J with $\tau^J \bar{\theta}_k \geq B_q \Delta_k^{\frac{q-2}{q-1}}$. At this trial, $\theta_{k,J} = B_q \Delta_k^{\frac{q-2}{q-1}}$ and $\alpha_{k,J} = (1 + B_q \Delta_k^{\frac{q-2}{q-1}})^{-1}$. If $w_{k,J} = 0$, then $x_{k,J} = x_k$ and $r_{k+1,J} = 0$, so (3.3) holds trivially. Otherwise, $w_{k,J} \neq 0$. Applying Lemma 3.2 to this trial gives $\|w_{k,J}\|_{x_k}^* \leq \kappa_\omega \Delta_k$, and hence

$$\theta_{k,J} = B_q \Delta_k^{\frac{q-2}{q-1}} = (9L_{p,\nu})^{\frac{1}{q-1}} (\kappa_\omega \Delta_k)^{\frac{q-2}{q-1}} \geq (9L_{p,\nu})^{\frac{1}{q-1}} \|w_{k,J}\|_{x_k}^{*\frac{q-2}{q-1}}.$$

Thus, Proposition C.1 gives (3.3). When $p = 2$ and $\nu = 0$, we have $q = 2$ and $\theta_{k,J} = B_2 = 9L_{2,0}$, so the same argument applies. In either case, backtracking terminates no later than trial J. □

C.5 Proof of Proposition 3.5

Lemma C.5 (Uniform positive definiteness and comparison of local dual norms)**.** Suppose that Assumption 3.1 holds with $q > 2$. Then, there exist $\mu, L > 0$ such that

$$\mu I \preceq \nabla^2 f(x) \preceq LI, \qquad x \in \mathcal{X}. \tag{C.13}$$

Moreover, for any $x, y \in \mathcal{X}$ and $r \in \mathbb{R}^n$,

$$\|r\|_x^{*2} \geq \frac{\mu}{L} \|r\|_y^{*2}. \tag{C.14}$$

Proof. If $p = 2$ and $\nu > 0$, then Assumption 3.1 implies that $\nabla^2 f$ is continuous, while $\nabla^2 f$ is also continuous when $p = 3$. Moreover, $\nabla^2 f(x) \succ 0$ holds by the assumption on f. Therefore, by the compactness of $\mathcal{X}$, there exist $\mu, L > 0$ satisfying (C.13). Hence,

$$\|r\|_x^{*2} = \langle \nabla^2 f(x)^{-1} r, r \rangle \geq \frac{1}{L} \|r\|^2 \geq \frac{\mu}{L} \langle \nabla^2 f(y)^{-1} r, r \rangle = \frac{\mu}{L} \|r\|_y^{*2}.$$

□

Proof of Proposition 3.5. If the method terminates at initialization, there is no update to consider. Otherwise, $g_0 > 0$ implies $\Delta_0 > 0$, and $\Delta_k \geq \rho^k \Delta_0 > 0$ for every generated iterate. By Theorem 3.3, every RBNFW update satisfies (3.5). By (3.5) and (C.14),

$$f(x_k) - f(x_{k+1}) \overset{(3.5)}{\geq} \frac{c_p}{B_q} \frac{\|r_{k+1}\|_{x_k}^{*2}}{\Delta_k^{\frac{q-2}{q-1}}} - \eta_k \overset{(C.14)}{\geq} \frac{c_p \mu}{B_q L} \frac{\|r_{k+1}\|_{x_{k+1}}^{*2}}{\Delta_k^{\frac{q-2}{q-1}}} - \eta_k. \tag{C.15}$$

Moreover, the accuracy (3.2) gives

$$\eta_k = \frac{\omega \Delta_k^2}{1 + B_q \Delta_k^{\frac{q-2}{q-1}}} \leq \frac{\omega}{B_q} \Delta_k^{2 - \frac{q-2}{q-1}} = \frac{\omega}{B_q} \Delta_k^{\frac{q}{q-1}}. \tag{C.16}$$

First, suppose that $\|r_{k+1}\|_{x_{k+1}}^* \geq \rho \Delta_k$. Then, $\Delta_{k+1} = \|r_{k+1}\|_{x_{k+1}}^*$. Since $\Delta_{k+1} \geq \rho \Delta_k$ and $\frac{q}{q-1} = 2 - \frac{q-2}{q-1}$, we have

$$\Delta_k^{\frac{q}{q-1}} = \frac{\Delta_k^2}{\Delta_k^{\frac{q-2}{q-1}}} \leq \rho^{-2} \frac{\Delta_{k+1}^2}{\Delta_k^{\frac{q-2}{q-1}}}, \qquad \Delta_{k+1}^{\frac{q}{q-1}} = \frac{\Delta_{k+1}^2}{\Delta_{k+1}^{\frac{q-2}{q-1}}} \leq \rho^{-\frac{q-2}{q-1}} \frac{\Delta_{k+1}^2}{\Delta_k^{\frac{q-2}{q-1}}}.$$

Hence, by (C.15) and (C.16),

$$\begin{aligned}
\Phi(x_k) - \Phi(x_{k+1}) &= f(x_k) - f(x_{k+1}) + \chi \left(\Delta_k^{\frac{q}{q-1}} - \Delta_{k+1}^{\frac{q}{q-1}} \right) \\
&\geq \frac{c_p \mu}{B_q L} \frac{\Delta_{k+1}^2}{\Delta_k^{\frac{q-2}{q-1}}} - \frac{\omega}{B_q} \Delta_k^{\frac{q}{q-1}} - \chi \Delta_{k+1}^{\frac{q}{q-1}} \\
&\geq \left(\frac{c_p \mu}{B_q L} - \frac{\omega}{B_q \rho^2} - \chi \rho^{-\frac{q-2}{q-1}} \right) \frac{\Delta_{k+1}^2}{\Delta_k^{\frac{q-2}{q-1}}}.
\end{aligned}$$

Next, suppose that $\|r_{k+1}\|^*_{x_{k+1}} < \rho\Delta_k$. Then, $\Delta_{k+1} = \rho\Delta_k$. Moreover,

$$\Delta_k^{\frac{q}{q-1}} = \frac{1}{\rho^2}\frac{\rho^2\Delta_k^2}{\Delta_k^{\frac{q-2}{q-1}}} = \frac{1}{\rho^2}\frac{\Delta_{k+1}^2}{\Delta_k^{\frac{q-2}{q-1}}}.$$

Therefore, by (C.15) and (C.16),

$$\begin{aligned}\Phi(x_k) - \Phi(x_{k+1}) &= f(x_k) - f(x_{k+1}) + \chi\left(\Delta_k^{\frac{q}{q-1}} - \Delta_{k+1}^{\frac{q}{q-1}}\right)\\ &\geq -\frac{\omega}{B_q}\Delta_k^{\frac{q}{q-1}} + \chi\left(1-\rho^{\frac{q}{q-1}}\right)\Delta_k^{\frac{q}{q-1}}\\ &= \frac{\chi\left(1-\rho^{\frac{q}{q-1}}\right) - \omega/B_q}{\rho^2}\frac{\Delta_{k+1}^2}{\Delta_k^{\frac{q-2}{q-1}}}.\end{aligned}$$

Condition (3.7) is equivalent to the interval

$$\frac{\omega}{B_q(1-\rho^{\frac{q}{q-1}})} < \chi < \frac{\rho^{\frac{q-2}{q-1}}}{B_q}\left(\frac{c_p\mu}{L} - \frac{\omega}{\rho^2}\right) \tag{C.17}$$

being nonempty. Choose χ in this interval. Multiplying the upper bound in (C.17) by $\rho^{-\frac{q-2}{q-1}}$ and rearranging gives

$$\frac{c_p\mu}{B_qL} - \frac{\omega}{B_q\rho^2} - \chi\rho^{-\frac{q-2}{q-1}} > 0.$$

Rearranging the lower bound in (C.17) gives

$$\frac{\chi\left(1-\rho^{\frac{q}{q-1}}\right) - \omega/B_q}{\rho^2} > 0.$$

Hence,

$$c_\chi \coloneqq \min\left\{\frac{c_p\mu}{B_qL} - \frac{\omega}{B_q\rho^2} - \chi\rho^{-\frac{q-2}{q-1}}, \frac{\chi(1-\rho^{\frac{q}{q-1}}) - \omega/B_q}{\rho^2}\right\} > 0, \tag{C.18}$$

and (3.8) holds in both cases. □

C.6 Proof of Theorem 3.6

Lemma C.6 (Objective error bound in terms of the residual)**.** Suppose that f is μ-strongly convex and L-smooth on $\mathcal{X}$. Let $x \in \mathcal{X}$, $s \in \mathbb{R}^n$, and $\eta \geq 0$ satisfy

$$\max_{v\in\mathcal{X}}\langle s, v - x\rangle \leq \eta, \tag{C.19}$$

and let $r \coloneqq \nabla f(x) + s$. Then,

$$f(x) - f(x^\star) \leq \frac{L}{2\mu}\|r\|_x^{*2} + \eta.$$

Proof. By the μ-strong convexity of f,

$$\begin{aligned}f(x) - f(x^\star) &\leq \langle\nabla f(x), x - x^\star\rangle - \frac{\mu}{2}\|x - x^\star\|^2\\ &= \langle r, x - x^\star\rangle + \langle s, x^\star - x\rangle - \frac{\mu}{2}\|x - x^\star\|^2\\ &\overset{(a)}{\leq} \sqrt{L}\|r\|_x^*\|x - x^\star\| + \eta - \frac{\mu}{2}\|x - x^\star\|^2\\ &\overset{(b)}{\leq} \frac{L}{2\mu}\|r\|_x^{*2} + \eta.\end{aligned}$$

In (a), we used (C.19), the Cauchy–Schwarz inequality, and $\nabla^2 f(x) \preceq LI$ to obtain

$$\langle r, x - x^\star \rangle \le \|r\|_x^* \|x - x^\star\|_x \le \sqrt{L}\|r\|_x^* \|x - x^\star\|.$$

In (b), we used Young's inequality

$$\sqrt{L}\|r\|_x^* \|x - x^\star\| \le \frac{L}{2\mu}\|r\|_x^{*2} + \frac{\mu}{2}\|x - x^\star\|^2.$$

□

Proof of Theorem 3.6. If $g_0 = 0$, the method terminates immediately and the bound at $k = 0$ is trivial. Otherwise, $\Delta_0 > 0$, and Corollary 3.4 guarantees that every outer update is well defined. Let $i \ge 0$ be an index for which the iterate is generated. For $i = 0$, set $s_{-1} = 0$ and $\eta_{-1} = 0$, so (C.19) holds immediately. For $i \ge 1$, (2.2) at iteration $i - 1$ yields (C.19) with $x = x_i$, $s = s_{i-1}$, and $\eta = \eta_{i-1}$. Therefore, applying Lemma C.6 with $r = r_i$ gives

$$f(x_i) - f(x^\star) \le \frac{L}{2\mu}\|r_i\|_{x_i}^{*2} + \eta_{i-1}.$$

For $i \ge 1$, the accuracy and $\Delta_i \ge \rho\Delta_{i-1}$ imply

$$\eta_{i-1} = \frac{\omega\Delta_{i-1}^2}{1 + B_q\Delta_{i-1}^{\frac{q-2}{q-1}}} \le \omega\Delta_{i-1}^2 \le \omega\rho^{-2}\Delta_i^2.$$

The same bound holds for $i = 0$ because $\eta_{-1} = 0$. Therefore,

$$f(x_i) - f(x^\star) \le \left(\frac{L}{2\mu} + \omega\rho^{-2}\right)\Delta_i^2. \tag{C.20}$$

We next use (3.8). Since $\chi\Delta_i^{\frac{q}{q-1}} \le \Phi(x_i) \le \Phi(x_0)$, we have $\Delta_i \le \left(\frac{\Phi(x_0)}{\chi}\right)^{\frac{q-1}{q}}$. Hence, (C.20) yields

$$f(x_i) - f(x^\star) \le \left(\frac{L}{2\mu} + \omega\rho^{-2}\right)\Delta_i^2 \le \left(\frac{L}{2\mu} + \omega\rho^{-2}\right)\left(\frac{\Phi(x_0)}{\chi}\right)^{\frac{q-2}{q}}\Delta_i^{\frac{q}{q-1}}.$$

Therefore,

$$\Phi(x_i) = f(x_i) - f(x^\star) + \chi\Delta_i^{\frac{q}{q-1}} \le \left(\chi + \left(\frac{L}{2\mu} + \omega\rho^{-2}\right)\left(\frac{\Phi(x_0)}{\chi}\right)^{\frac{q-2}{q}}\right)\Delta_i^{\frac{q}{q-1}}. \tag{C.21}$$

Now let i be a nonterminal index, so x_{i+1} is generated and $\Delta_i, \Delta_{i+1} > 0$. Since $\Delta_{i+1} \ge \rho\Delta_i$ and $\frac{q}{q-1} = 2 - \frac{q-2}{q-1}$,

$$\frac{\Delta_{i+1}^2}{\Delta_i^{\frac{q-2}{q-1}}} = \Delta_{i+1}^{2-\frac{q-2}{q-1}}\left(\frac{\Delta_{i+1}}{\Delta_i}\right)^{\frac{q-2}{q-1}} = \Delta_{i+1}^{\frac{q}{q-1}}\left(\frac{\Delta_{i+1}}{\Delta_i}\right)^{\frac{q-2}{q-1}} \ge \rho^{\frac{q-2}{q-1}}\Delta_{i+1}^{\frac{q}{q-1}}.$$

Moreover, (C.21) gives

$$\Delta_{i+1}^{\frac{q}{q-1}} \ge \frac{\Phi(x_{i+1})}{\chi + \left(\frac{L}{2\mu} + \omega\rho^{-2}\right)\left(\frac{\Phi(x_0)}{\chi}\right)^{\frac{q-2}{q}}}.$$

Consequently, (3.8) implies

$$\Phi(x_i) - \Phi(x_{i+1}) \ge \frac{c_\chi\rho^{\frac{q-2}{q-1}}}{\chi + \left(\frac{L}{2\mu} + \omega\rho^{-2}\right)\left(\frac{\Phi(x_0)}{\chi}\right)^{\frac{q-2}{q}}}\Phi(x_{i+1}).$$

Equivalently,

$$\Phi(x_{i+1}) \le \left(1 + \frac{c_\chi\rho^{\frac{q-2}{q-1}}}{\chi + \left(\frac{L}{2\mu} + \omega\rho^{-2}\right)\left(\frac{\Phi(x_0)}{\chi}\right)^{\frac{q-2}{q}}}\right)^{-1}\Phi(x_i).$$

Iterating this inequality for $i = 0, 1, \dots, k-1$ yields the upper bound in (3.9), while the lower bound follows from $f(x_k) - f(x^\star) \le \Phi(x_k)$. □

D PROOFS OF SECTION 3.2

After a finite number of iterations, we switch to the full-step regime by setting $\alpha_k = 1$ and solving each subproblem until the stopping criterion in (3.10) is satisfied. The local analysis applies to RBNFW after switching.

D.1 PROOF OF LEMMA 3.7

Proof of Lemma 3.7. Let $M_k^* := f(x_k) - \Psi_1(\widehat{x}_{k+1}; x_k)$. Since $g_k > 0$, there exists $v \in \mathcal{X}$ such that $\langle \nabla f(x_k), v - x_k \rangle < 0$. Hence, for all $0 < \gamma < \min\left\{1, \frac{-2\langle \nabla f(x_k), v - x_k\rangle}{\|v - x_k\|_{x_k}^2}\right\}$, the point $x_k + \gamma(v - x_k)$ is feasible and

$$\Psi_1(x_k + \gamma(v - x_k); x_k) - f(x_k) = \gamma\langle \nabla f(x_k), v - x_k\rangle + \frac{\gamma^2}{2}\|v - x_k\|_{x_k}^2 < 0,$$

and therefore $M_k^* > 0$ due to $\Psi_1(\widehat{x}_{k+1}; x_k) \le \Psi_1(x_k + \gamma(v - x_k); x_k) < f(x_k)$.

Since the inner updates use exact line minimization,

$$\Psi_1(u_{t+1}; x_k) \le \Psi_1(u_t; x_k) \le \Psi_1(u_0; x_k) = f(x_k),$$

and hence $0 \le M_t \le M_k^*$. The convexity of $\Psi_1(\cdot; x_k)$ gives

$$0 \le M_k^* - M_t = \Psi_1(u_t; x_k) - \Psi_1(\widehat{x}_{k+1}; x_k) \le G_t.$$

Fix any $a > 0$ and suppose, for contradiction, that $\ell_k(a)$ is infinite, so the inner updates continue indefinitely. For FW, Proposition 2.1 guarantees finite termination for every positive FW-gap tolerance. For away-step FW, the same follows from the inner-loop argument in the proof of Corollary 3.4 with $\alpha = 1$. Since the update rules do not depend on the stopping tolerance, applying these arguments to the same inner sequence with positive tolerances tending to zero yields inner indices $\{t_j\}$ such that $G_{t_j} \to 0$. Consequently, $M_{t_j} \to M_k^* > 0$, and, for any $a > 0$,

$$\left(\frac{M_{t_j}}{\|\nabla f(x_k)\|}\right)^a \to \left(\frac{M_k^*}{\|\nabla f(x_k)\|}\right)^a > 0.$$

Therefore, the condition in (3.10) holds for all sufficiently large j, a contradiction. Hence, $\ell_k(a)$ is finite. Furthermore, by the μ-strong convexity of $\Psi_1(\cdot; x_k)$ and (3.10),

$$\frac{\mu}{2}\|u_{\ell_k(a)} - \widehat{x}_{k+1}\|^2 \le G_{\ell_k(a)} \le \left(\frac{M_{\ell_k(a)}}{\|\nabla f(x_k)\|}\right)^a.$$

On the other hand,

$$M_{\ell_k(a)} \le M_k^* = -\langle \nabla f(x_k), \widehat{x}_{k+1} - x_k\rangle - \frac{1}{2}\|\widehat{x}_{k+1} - x_k\|_{x_k}^2 \le \|\nabla f(x_k)\|\|\widehat{x}_{k+1} - x_k\|,$$

where the last inequality uses the Cauchy–Schwarz inequality and $-\frac{1}{2}\|\widehat{x}_{k+1} - x_k\|_{x_k}^2 < 0$. Combining the last two displays yields (3.11) and (3.12). □

D.2 PROOF OF THEOREM 3.8

Lemma D.1 (Vanishing of the FW gap)**.** Let $\{x_k\} \subset \mathcal{X}$ satisfy $f(x_k) - f(x^\star) \to 0$. Then, $g_k \to 0$.

Proof. By the strong convexity of f, $\frac{\mu}{2}\|x_k - x^\star\|^2 \le f(x_k) - f(x^\star) \to 0$. Hence, $x_k \to x^\star$. Since $\mathcal{X}$ is compact and ∇f is continuous, the FW-gap function $x \mapsto \max_{v \in \mathcal{X}}\langle \nabla f(x), x - v\rangle$ is continuous. Moreover, the optimality of $x^\star$ implies that its FW gap is zero. Therefore, $g_k \to 0$. □

Proof of Theorem 3.8. If the switching criterion in (3.13) were never satisfied, RBNFW with (3.2) would be used at every iteration. Then, Theorem 3.6 and Lemma D.1 would give $g_k \to 0$, contradicting the failure of the switching criterion. Therefore, K_2 is finite.

If $g_j = 0$ for some $j \ge K_2$, then the strong convexity of f gives $x_j = x^\star$, so the sequence reaches the solution in finitely many iterations. Hence, it remains to consider the case $g_k > 0$ for every

$k \geq K_2$. Let $k \geq K_2$, define $e_k := x_k - x^\star$, and let $\widehat{x}_{k+1}$ be the exact minimizer of $\Psi_1(\cdot; x_k)$ over $\mathcal{X}$. By the optimality conditions of $\widehat{x}_{k+1}$ and $x^\star$,

$$\|\widehat{x}_{k+1} - x^\star\|_{x_k}^2 \leq \left\langle \nabla^2 f(x_k) e_k - (\nabla f(x_k) - \nabla f(x^\star)), \widehat{x}_{k+1} - x^\star \right\rangle.$$

By the Cauchy–Schwarz inequality and the Hölder continuity of the Hessian,

$$\begin{aligned}\|\widehat{x}_{k+1} - x^\star\|_{x_k} &\leq \left\| \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^\star + t e_k)\right) e_k \, dt \right\|_{x_k}^* \\ &\leq L_{2,\nu} \|e_k\|_{x_k}^{1+\nu} \int_0^1 (1-t)^\nu \, dt = \frac{L_{2,\nu}}{1+\nu} \|e_k\|_{x_k}^{1+\nu}.\end{aligned}$$

Therefore,

$$\|\widehat{x}_{k+1} - x^\star\| \leq \frac{L_{2,\nu} L^{\frac{1+\nu}{2}}}{(1+\nu)\sqrt{\mu}} \|e_k\|^{1+\nu}. \tag{D.1}$$

Using (D.1), $\|e_k\| \leq D/\sqrt{\mu}$, and the triangle inequality,

$$\|\widehat{x}_{k+1} - x_k\| \leq \|\widehat{x}_{k+1} - x^\star\| + \|x^\star - x_k\| \leq \left(1 + \frac{L_{2,\nu} L^{\frac{1+\nu}{2}} D^\nu}{(1+\nu)\mu^{\frac{1+\nu}{2}}}\right) \|e_k\|.$$

Combining this inequality with (3.12) with $a = 2(1+\nu)$ gives

$$\|x_{k+1} - \widehat{x}_{k+1}\| \leq \sqrt{\frac{2}{\mu}} \left(1 + \frac{L_{2,\nu} L^{\frac{1+\nu}{2}} D^\nu}{(1+\nu)\mu^{\frac{1+\nu}{2}}}\right)^{1+\nu} \|e_k\|^{1+\nu}.$$

Hence, by (D.1) and the triangle inequality,

$$\|x_{k+1} - x^\star\| \leq \|x_{k+1} - \widehat{x}_{k+1}\| + \|\widehat{x}_{k+1} - x^\star\| \leq C_2 \|e_k\|^{1+\nu}.$$

This proves (3.14).

By (3.13), strong convexity, and convexity,

$$\frac{\mu}{2}\|e_{K_2}\|^2 \leq f(x_{K_2}) - f(x^\star) \leq g_{K_2} \leq \frac{\mu}{2}(2C_2)^{-\frac{2}{\nu}}.$$

Therefore, the switching condition implies $C_2\|e_{K_2}\|^\nu \leq \frac{1}{2}$. Moreover, if $C_2\|e_k\|^\nu \leq 1/2$, then

$$\|e_{k+1}\| \leq C_2 \|e_k\|^{1+\nu} \leq \frac{1}{2}\|e_k\|, \qquad C_2\|e_{k+1}\|^\nu \leq 2^{-\nu} C_2 \|e_k\|^\nu \leq \frac{1}{2}.$$

Thus, by induction, $x_k \to x^\star$ and $\|e_{k+1}\| \leq \frac{1}{2}\|e_k\|$ for every $k \geq K_2$.

Since $\alpha_k = 1$, the definition of s_k, Lemma C.3, $\nabla^2 f(x_k) \preceq LI$, and the preceding contraction give

$$\|\nabla f(x_{k+1}) + s_k\|_{x_k}^* \leq \frac{L_{2,\nu}}{1+\nu} \|x_{k+1} - x_k\|_{x_k}^{1+\nu} \leq \frac{L_{2,\nu}}{1+\nu} \left(\frac{3\sqrt{L}}{2}\right)^{1+\nu} \|e_k\|^{1+\nu}.$$

Moreover, (3.11) and the bound on $\|\widehat{x}_{k+1} - x_k\|$ established above imply

$$G_{\ell_k(2(1+\nu))} \leq \left(1 + \frac{L_{2,\nu} L^{(1+\nu)/2} D^\nu}{(1+\nu)\mu^{(1+\nu)/2}}\right)^{2(1+\nu)} \|e_k\|^{2(1+\nu)}.$$

Since $\nabla \Psi_1(x_{k+1}; x_k) = -s_k$, we have $\max_{v \in \mathcal{X}} \langle s_k, v - x_{k+1} \rangle = G_{\ell_k(2(1+\nu))}$. Applying Lemma C.6 with $x = x_{k+1}$, $s = s_k$, and $\eta = G_{\ell_k(2(1+\nu))}$, and using (C.14) in the form $\|\nabla f(x_{k+1}) + s_k\|_{x_{k+1}}^{*2} \leq \frac{L}{\mu}\|\nabla f(x_{k+1}) + s_k\|_{x_k}^{*2}$ together with $\frac{\mu}{2}\|e_k\|^2 \leq f(x_k) - f(x^\star)$, we obtain, for some constant $\widetilde{C}_2 > 0$ independent of k,

$$f(x_{k+1}) - f(x^\star) \leq \frac{L}{2\mu}\|\nabla f(x_{k+1}) + s_k\|_{x_{k+1}}^{*2} + G_{\ell_k(2(1+\nu))} \leq \widetilde{C}_2 \left(f(x_k) - f(x^\star)\right)^{1+\nu}.$$

This proves (3.15) and completes the proof. □

D.3 Proof of Theorem 3.9

Proof of Theorem 3.9. If the switching criterion in (3.17) were never satisfied, RBNFW with (3.2) would be used at every iteration. Then, Theorem 3.6 and Lemma D.1 would give $g_k \to 0$, contradicting the failure of the switching criterion. Therefore, K_3 is finite.

If $g_j = 0$ for some $j \geq K_3$, then the strong convexity of f gives $x_j = x^\star$, so the sequence reaches the solution in finitely many iterations. Hence, it remains to consider the case $g_k > 0$ for every $k \geq K_3$. Let $k \geq K_3$, define $e_k \coloneqq x_k - x^\star$, and let $\widehat{x}_{k+1}$ be the exact minimizer of $\Psi_1(\cdot; x_k)$ over $\mathcal{X}$. By the optimality conditions of $\widehat{x}_{k+1}$ and $x^\star$,

$$\left\langle \nabla f(x_k) + \nabla^2 f(x_k)(\widehat{x}_{k+1} - x_k), x^\star - \widehat{x}_{k+1} \right\rangle \geq 0, \qquad \langle \nabla f(x^\star), \widehat{x}_{k+1} - x^\star \rangle \geq 0.$$

Therefore,

$$\|\widehat{x}_{k+1} - x^\star\|_{x_k}^2 \leq \left\langle \nabla^2 f(x_k) e_k - (\nabla f(x_k) - \nabla f(x^\star)), \widehat{x}_{k+1} - x^\star \right\rangle.$$

By the Cauchy–Schwarz inequality,

$$\|\widehat{x}_{k+1} - x^\star\|_{x_k} \leq \|\nabla^2 f(x_k) e_k - (\nabla f(x_k) - \nabla f(x^\star))\|_{x_k}^*.$$

Using $\mu I \preceq \nabla^2 f(x_k)$ and (3.16), we obtain

$$\begin{aligned}
\|\widehat{x}_{k+1} - x^\star\| &\leq \frac{1}{\sqrt{\mu}} \|\widehat{x}_{k+1} - x^\star\|_{x_k} \\
&\leq \frac{1}{\sqrt{\mu}} \|\nabla^2 f(x_k) e_k - (\nabla f(x_k) - \nabla f(x^\star))\|_{x_k}^* \\
&\leq \frac{1}{\mu} \left\| \nabla^2 f(x_k) e_k - (\nabla f(x_k) - \nabla f(x^\star)) \right\| \\
&= \frac{1}{\mu} \left\| \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^\star + t e_k) \right) e_k \, dt \right\| \\
&\leq \frac{1}{\mu} \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^\star + t e_k)\|_2 \|e_k\| \, dt \\
&\leq \frac{M}{\mu} \|e_k\|^2 \int_0^1 (1 - t) \, dt = \frac{M}{2\mu} \|e_k\|^2.
\end{aligned} \tag{D.2}$$

Using (D.2), $\|e_k\| \leq D/\sqrt{\mu}$, and the triangle inequality,

$$\|\widehat{x}_{k+1} - x_k\| \leq \|e_k\| + \|\widehat{x}_{k+1} - x^\star\| \leq \left(1 + \frac{M}{2\mu} \|e_k\|\right) \|e_k\| \leq \left(1 + \frac{MD}{2\mu^{3/2}}\right) \|e_k\|.$$

Combining this inequality with (3.12) with $a = 4$ gives $\|x_{k+1} - \widehat{x}_{k+1}\| \leq \sqrt{\frac{2}{\mu}} \left(1 + \frac{MD}{2\mu^{3/2}}\right)^2 \|e_k\|^2$. Hence, by (D.2) and the triangle inequality,

$$\|x_{k+1} - x^\star\| \leq \|x_{k+1} - \widehat{x}_{k+1}\| + \|\widehat{x}_{k+1} - x^\star\| \leq C_3 \|e_k\|^2.$$

This proves (3.18).

By (3.17), strong convexity, and convexity,

$$\frac{\mu}{2} \|e_{K_3}\|^2 \leq f(x_{K_3}) - f(x^\star) \leq g_{K_3} \leq \frac{\mu}{8C_3^2}.$$

Therefore, $C_3 \|e_{K_3}\| \leq \frac{1}{2}$. If the same inequality holds at iteration k, then (3.18) gives $\|e_{k+1}\| \leq \frac{1}{2}\|e_k\|$ and $C_3\|e_{k+1}\| \leq \frac{1}{4}$. Thus, by induction, $x_k \to x^\star$ and $\|e_{k+1}\| \leq \frac{1}{2}\|e_k\|$ for every $k \geq K_3$.

Since $\alpha_k = 1$, the definition of s_k, (3.16), and the preceding contraction give

$$\begin{aligned}
\|\nabla f(x_{k+1}) + s_k\|_{x_{k+1}}^* &\leq \frac{1}{\sqrt{\mu}} \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k)\| \\
&\leq \frac{M}{2\sqrt{\mu}} \|x_{k+1} - x_k\|^2 \leq \frac{9M}{8\sqrt{\mu}} \|e_k\|^2.
\end{aligned}$$

Moreover, (3.11) and the bound on $\|\widehat{x}_{k+1} - x_k\|$ established above imply $G_{\ell_k(4)} \leq \left(1 + \frac{MD}{2\mu^{3/2}}\right)^4 \|e_k\|^4$. Since $\nabla\Psi_1(x_{k+1}; x_k) = -s_k$, we have $\max_{v\in\mathcal{X}}\langle s_k, v - x_{k+1}\rangle = G_{\ell_k(4)}$. Applying Lemma C.6 with $x = x_{k+1}$, $s = s_k$, and $\eta = G_{\ell_k(4)}$, and using the preceding bounds together with $\frac{\mu}{2}\|e_k\|^2 \leq f(x_k) - f(x^\star)$, we obtain, for some constant $\widetilde{C}_3 > 0$ independent of k,

$$f(x_{k+1}) - f(x^\star) \leq \frac{L}{2\mu}\|\nabla f(x_{k+1}) + s_k\|^{*2}_{x_{k+1}} + G_{\ell_k(4)} \leq \widetilde{C}_3\left(f(x_k) - f(x^\star)\right)^2.$$

This proves (3.19) and completes the proof. □

D.4 Total LMO Complexity

Finally, we summarize the total LMO complexity of our methods and several related methods. For the global bounds, we use the global stepsize and inner accuracy rules throughout under the assumptions of Theorem 3.6. By (3.2), (C.20), and $\Phi(x_k) \leq \Phi(x_0)$, we have $\eta_k \geq c\epsilon$ and $\alpha_{k,j} \geq \underline{\alpha}$ for all trials whenever $f(x_k) - f(x^\star) > \epsilon$, with $c, \underline{\alpha} > 0$ independent of ϵ. Since g_k is bounded and $|\widehat{\mathcal{S}}_k| \leq |\operatorname{Vert}\mathcal{X}| < \infty$ in the polytope case, Propositions 2.1 and 2.2 give $\mathcal{O}(\epsilon^{-1})$ and $\mathcal{O}(\log\epsilon^{-1})$ LMO calls per inner solve, respectively. For the accepted trial index j_k in Algorithm 4, acceptance at the root value, $\Delta_k \geq \rho\Delta_{k-1}$, and the update $\theta_{k+1} = \theta_k/\tau$ give $j_k \leq 2 + \frac{q-2}{q-1}\log_\tau(\rho^{-1}) + \log_\tau(\bar{\theta}_{k+1}/\bar{\theta}_k)$ for $k \geq 1$. Since $\bar{\theta}_k$ is bounded above and j_0 is bounded independently of ϵ, telescoping gives $\mathcal{O}(1)$ inner solves per outer iteration on average; the root Newton rule uses one. Combining these estimates with the $\mathcal{O}(\log\epsilon^{-1})$ outer iteration bound of Theorem 3.6 yields global total LMO complexities $\mathcal{O}(\epsilon^{-1}\log\epsilon^{-1})$ and $\mathcal{O}((\log\epsilon^{-1})^2)$ for Algorithms 1 and 3, respectively. These bounds hold with problem and algorithmic parameters fixed and include initialization, outer FW-gap evaluations, and rejected trials.

For the local bounds, we set $a = 2(1+\nu)$ for $p = 2$ and $a = 4$ for $p = 3$. Writing $M_k^* := f(x_k) - \Psi_1(\widehat{x}_{k+1}; x_k)$, comparison with the feasible point $x_k + (\mu/L)(x^\star - x_k)$ gives $M_k^* \geq (\mu/L)(f(x_k) - f(x^\star))$ by strong convexity and $\nabla^2 f(x_k) \preceq LI$. Define the analysis-only tolerance $\bar{\eta}_k := \min\left\{\frac{M_k^*}{2}, \left(\frac{M_k^*}{2\|\nabla f(x_k)\|}\right)^a\right\}$. Since $M_k^* - M_t \leq G_t$, the inequality $G_t \leq \bar{\eta}_k$ implies $M_t \geq M_k^*/2$ and hence the stopping condition (3.10). Thus, the local inner loop terminates no later than the same inner method run with fixed tolerance $\bar{\eta}_k$. Since ∇f is bounded on $\mathcal{X}$, there exists $c > 0$, independent of k, such that $\bar{\eta}_k \geq c(f(x_k) - f(x^\star))^a$ whenever $0 < f(x_k) - f(x^\star) \leq 1$. Thus, for $0 < f(x_k) - f(x^\star) \leq 1/e$, Proposition 2.1 gives $\mathcal{O}((f(x_k) - f(x^\star))^{-a})$ LMO calls per FW inner solve, while the contraction estimate in the proof of Proposition 2.2 gives $\mathcal{O}(\log(f(x_k) - f(x^\star))^{-1})$ non-drop steps per AFW inner solve. Finite termination has an ϵ-independent cost, so assume that $f(x_k) - f(x^\star) > 0$ for all k. By Theorems 3.8 and 3.9, $f(x_k) - f(x^\star) \to 0$ and $f(x_{k+1}) - f(x^\star) \leq \widetilde{C}_p(f(x_k) - f(x^\star))^{a/2}$ after switching. Fix $l \in (1, a/2)$. There exists k_0 after switching, independent of ϵ, such that $0 < f(x_k) - f(x^\star) \leq 1/e$, $f(x_{k+1}) - f(x^\star) \leq (f(x_k) - f(x^\star))/2$, and $\log(f(x_{k+1}) - f(x^\star))^{-1} \geq l\log(f(x_k) - f(x^\star))^{-1}$ for all $k \geq k_0$. For sufficiently small $\epsilon > 0$, let $N := \min\{k \geq k_0 \mid f(x_k) - f(x^\star) \leq \epsilon\}$. Since $f(x_{N-1}) - f(x^\star) > \epsilon$, geometric summation yields

$$\sum_{k=k_0}^{N-1}(f(x_k) - f(x^\star))^{-a} \leq \frac{(f(x_{N-1}) - f(x^\star))^{-a}}{1 - 2^{-a}} = \mathcal{O}(\epsilon^{-a}),$$

$$\sum_{k=k_0}^{N-1}\log(f(x_k) - f(x^\star))^{-1} \leq \frac{l}{l-1}\log(f(x_{N-1}) - f(x^\star))^{-1} = \mathcal{O}(\log\epsilon^{-1}).$$

Since active sets are carried between solves, the total number of drop steps is at most the number of FW steps plus the active-set size at switching. Hence, the local total LMO complexities, including outer FW-gap evaluations, are $\mathcal{O}(\epsilon^{-a})$ and $\mathcal{O}(\log\epsilon^{-1})$ for Algorithms 1 and 3, respectively. Problem and algorithmic parameters are fixed, including $\nu > 0$ when $p = 2$. The finite LMO cost before k_0, including rejected trials, is independent of ϵ and absorbed in these bounds.

These bounds and those for related methods are summarized in Table 2. The local bounds are conservative worst-case guarantees based on sufficient inner accuracy conditions for fast local convergence. In the experiments of Section 4, our methods were competitive in LMO calls and often achieved the best wall-clock performance among the tested methods.

Table 2: Comparison of global and local total LMO complexity for attaining $f(x_k) - f(x^\star) \leq \epsilon$. All inner LMO calls are included. For the proposed methods, global bounds use the global accuracy and step-size rules throughout, including rejected trials, whereas local bounds count calls after switching. Our $\mathcal{C}^2$ bounds require $\nu > 0$. Here, $\nu_{\mathrm{N}} \in (1, 1.139)$. cvx: convex, scvx: strongly convex, self-conc.: self-concordant.

Algorithm	f	∇f Lip.	$\nabla^p f$	$\mathcal{X}$	global	local
FW	$\mathcal{C}^1$ cvx	✔	✘	cvx	$\mathcal{O}(\epsilon^{-1})$	—
CGS	$\mathcal{C}^1$ cvx	✔	✘	cvx	$\mathcal{O}(\epsilon^{-1})$	—
r-CGS	$\mathcal{C}^1$ scvx	✔	✘	cvx	$\mathcal{O}(\epsilon^{-1})$	—
NFW	$\mathcal{C}^3$ cvx	✘	self-conc.	cvx	—	$\mathcal{O}(\epsilon^{-\nu_{\mathrm{N}}})$
SOCGS	$\mathcal{C}^2$ scvx	✔	$\nabla^2 f$ Lip.	polytope	$\mathcal{O}((\log \epsilon^{-1})^2)$	$\mathcal{O}(\log \epsilon^{-1} \log\log \epsilon^{-1})$
Alg. 1	$\mathcal{C}^2$ scvx	✔	$\nabla^2 f$ Hölder	cvx	$\mathcal{O}(\epsilon^{-1} \log \epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2(1+\nu)})$
Alg. 3	$\mathcal{C}^2$ scvx	✔	$\nabla^2 f$ Hölder	polytope	$\mathcal{O}((\log \epsilon^{-1})^2)$	$\mathcal{O}(\log \epsilon^{-1})$
Alg. 1	$\mathcal{C}^3$ scvx	✔	$\nabla^3 f$ Hölder	cvx	$\mathcal{O}(\epsilon^{-1} \log \epsilon^{-1})$	$\mathcal{O}(\epsilon^{-4})$
Alg. 3	$\mathcal{C}^3$ scvx	✔	$\nabla^3 f$ Hölder	polytope	$\mathcal{O}((\log \epsilon^{-1})^2)$	$\mathcal{O}(\log \epsilon^{-1})$

E Additional Experimental Results

E.1 Matrix Sensing with Squared Loss

We generated the ground truth matrix $X_\star \in \mathcal{M}_{R_{\mathrm{Tr}}}$, and formed the deterministic observations $(y_{\mathrm{obs}})_i = \langle A_i, X_\star \rangle$, $i = 1, \ldots, m$. We chose orthonormal matrices $V_1, \ldots, V_\varphi \in \mathbb{S}^n$ satisfying $\mathrm{Tr}(V_j) = 0$, and extended them to an orthonormal basis of $\mathbb{S}^n$ with $W_1, \ldots, W_{m-\varphi}$ chosen orthogonal to every V_j. For the condition number Λ of the objective function, we set $A_j = \sqrt{\lambda_j} V_j$ and $A_{\varphi+\ell} = W_\ell$, where $\lambda_j = \Lambda^{j/\varphi}$. Rather than choosing their span at random, we constructed the V_j so that the squared Frobenius norm of the projection of $X_0 - X_\star$ onto their span equals $\varpi \|X_0 - X_\star\|_{\mathrm{F}}^2$, where $\varpi \in (0, 1)$ denotes the prescribed projection ratio. This experiment was designed to verify conditioning and convergence behavior.

We present additional numerical results for matrix sensing with squared loss. Figures 5 and 6 show the primal gap for $\varpi = 0.5$ and $\varpi = 0.2$, respectively. Our methods outperformed the baseline methods in both cases. In particular, RBNFW (`local3`) was the best-performing method in both cases. APG obtained the smallest primal gap but required much more time than our methods.

E.2 Ridge-Regularized Logistic Regression

We present additional numerical results for ridge-regularized logistic regression. Figures 7 to 11 show the primal gap for different datasets and constraint sets. In Figures 7 to 9, the RBNFW variants outperformed the baseline methods. In Figure 10, NFW also performed well. In Figure 11, the performance of the RBNFW variants in terms of computation time was nearly identical to that of SOCGS. Note that Figure 9 does not include the results for away-step FW, and Figure 10 does not include the results for CGS because they were very slow in these experiments.

E.3 Outer Iterations Reaching the Inner Iteration Limit

We capped inner solves at 1000 iterations, as in Carderera & Pokutta (2026, Section 4). We report the affected outer iterations in Table 3. When the cap was reached before the prescribed accuracy was attained, we accepted the last inner iterate; this relaxation is not covered by our convergence analysis. The prescribed accuracy depends on global problem constants and can be overly stringent in practice. RBNFW exhibited convergent behavior in the reported experiments and remained consistently stable, supporting its use in practice.

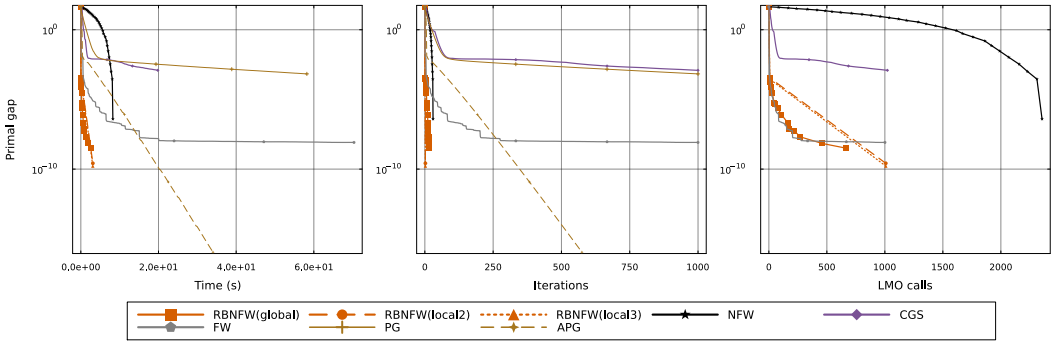


Figure 5: Primal gap for matrix sensing with $\varpi = 0.5$.

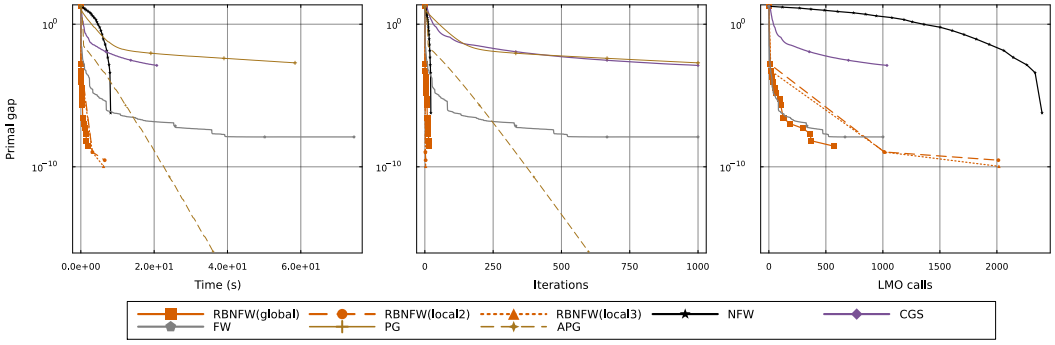


Figure 6: Primal gap for matrix sensing with $\varpi = 0.2$.

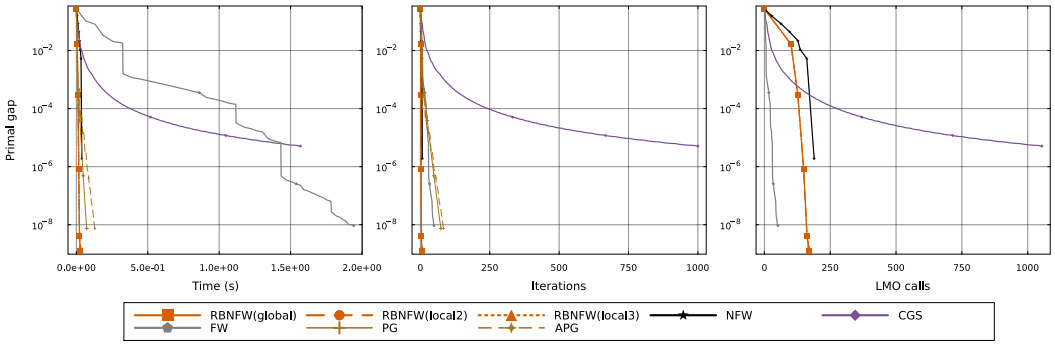


Figure 7: Primal gap for the a9a dataset with an ℓ_2-ball constraint.

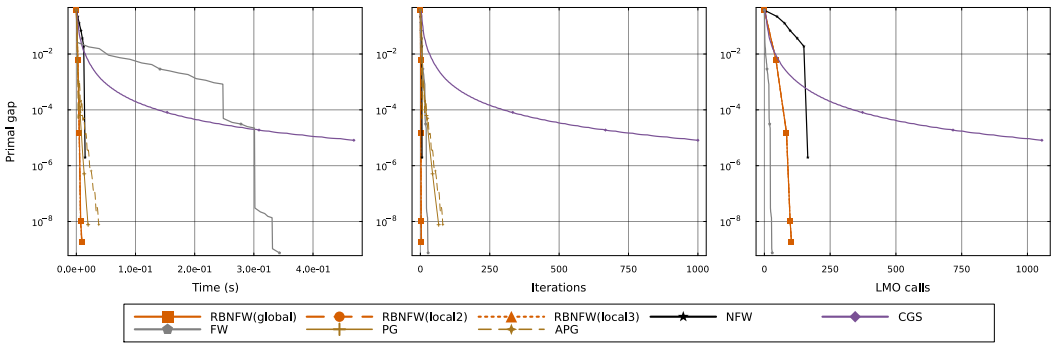


Figure 8: Primal gap for the mushrooms dataset with an ℓ_2-ball constraint.

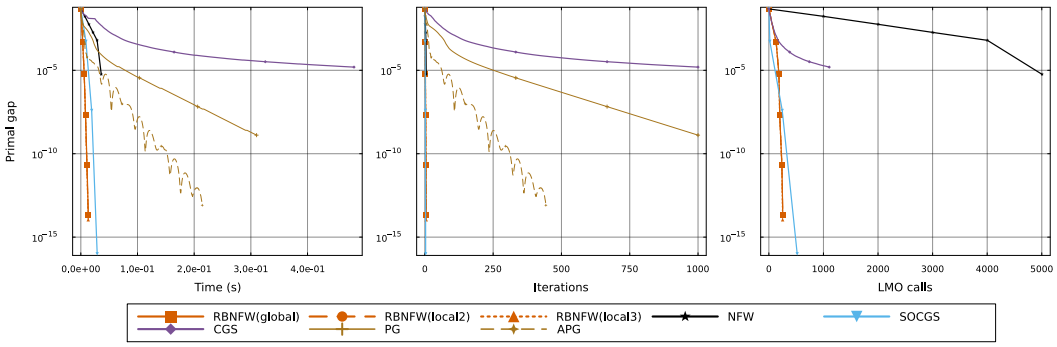


Figure 9: Primal gap for the mushrooms dataset with a sparse polytope constraint.

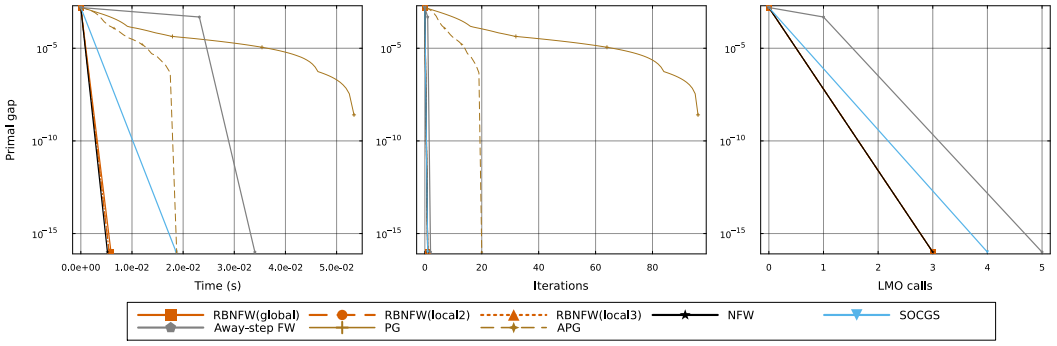


Figure 10: Primal gap for the phishing dataset with a sparse polytope constraint.

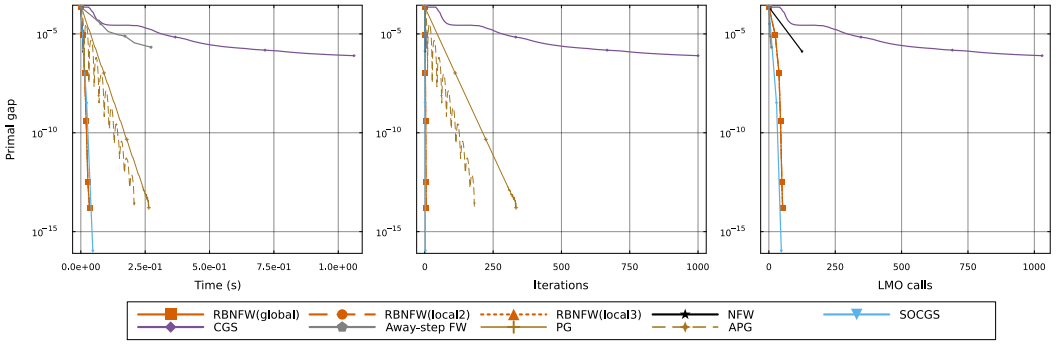


Figure 11: Primal gap for the w7a dataset with a sparse polytope constraint.

Table 3: Number of outer iterations in which at least one inner solve reached the maximum of 1000 inner iterations. Multiple capped inner solves within the same outer iteration are counted only once. An entry "–" indicates that the corresponding method was not included in the corresponding experiment.

	Matrix sensing			Logistic: ℓ_2-ball			Logistic: sparse polytope				
Method	$\varpi = 0.2$	$\varpi = 0.5$	$\varpi = 0.8$	covtype.binary	a9a	mushrooms	covtype.binary	a9a	mushrooms	phishing	w7a
RBNFW (`global`)	0	0	0	0	0	0	0	0	0	0	0
RBNFW (`local2`)	2	1	0	0	0	0	0	0	0	0	0
RBNFW (`local3`)	2	1	1	0	0	0	0	0	0	0	0
NFW	0	0	0	0	0	0	3	7	6	0	0
SOCGS	–	–	–	–	–	–	1	2	0	0	0
CGS	0	0	0	0	0	0	0	0	0	–	0