Perturbed Proximal Descent to Escape Saddle Points for Non-convex and Non-smooth Objective Functions

Zhishen Huang[1] and Stephen Becker[1 [0000−0002−1932−8159]]

Dept. of Applied Math., University of Colorado, Boulder, USA
{zhishen.huang,stephen.becker}@colorado.edu

Abstract. We consider the problem of finding local minimizers in non-convex and non-smooth optimization. Under the assumption of strict saddle points, result positive results have been derived for first-order methods. We present the first known results for the non-smooth case, which requires different analysis and a different algorithm. *This is the extended version of the paper that contains the proofs*

Keywords: Saddle-points · Proximal gradient descent · Non-smooth optimization.

1 Introduction

We consider the problem of finding approximate local minimizers of the problem

$$\text{minimize}_{\mathbf{x}\in\mathbb{R}^d}\ (\Phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})) \tag{1}$$

where $f(\mathbf{x})$ is not convex but smooth (and with full domain), and $g(\mathbf{x})$ is convex but not smooth. Many optimization problems in engineering, signal processing and machine learning can be cast in this framework, where f is a smooth loss function, and g is a non-smooth regularizer such as a norm. For example, our model captures regularized neural networks [11], where the regularization can induce sparsity as an alternative to dropout. In this paper, for simplicity we restrict our discussion to $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, where $\lambda \geq 0$ is a constant, but many of the results apply to more general choices of g. The first-order condition is $0 \in \nabla f(\mathbf{x}) + \partial g(\mathbf{x})$, and any $\mathbf{x}$ satisfying this condition is called a "stationary point" (see [2] for background on the subdifferential ∂g). All local minimizers are stationary points, but not vice-versa. We define a "saddle point" to be any stationary point where the Hessian is indefinite (and therefore not a local minimizer). This paper extends a recent line of work [13] to analyze when we can expect to find a local minimizer. It has been argued that in many machine learning problems, finding any local minimizer is often enough for good performance, but finding a saddle point is not useful [9].

The fact that g is non-smooth is crucially important, and it does more than just complicate the analysis, as it also requires a new algorithm. In the smooth

case, f is often minimized using gradient descent or an accelerated variant [16] with a fixed stepsize. Naïvely extending gradient descent to apply to (1) leads to subgradient descent with fixed-stepsize. Unfortunately, this method fails to converge as the example $d = 1, \lambda = 1$ and $f = 0$ shows [18] since for a generic choice of the initial point, the sequence is not Cauchy.

Instead of gradient descent, we use a perturbed version of proximal gradient descent. For a real-valued convex lower semi-continuous function g, define the "proximity" operator (or "prox" for short) as the map $\text{prox}_g(\mathbf{y}) = \text{argmin}_{\mathbf{x}}\, g(\mathbf{y}) + \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2$ (throughout the paper, for vectors we use $\|\cdot\|$ to denote the Euclidean norm). Equivalently, $\text{prox}_g = (I + \partial g)^{-1}$, and thus the first-order condition is equivalent to $\mathbf{x} = \text{prox}_{\eta g}[\mathbf{x} - \eta\nabla f(\mathbf{x})]$ for any $\eta > 0$. Proximal gradient descent is the iteration $\mathbf{x}_{t+1} = \text{prox}_{\eta g}[\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)]$, so it immediately follows that if the sequence converges, it converges to a stationary point. Convergence of the sequence is known to follow from mild assumptions on f and g, the stepsize η, and boundedness of the sequence $\{\mathbf{x}_t\}$ [1].

We define a *second-order stationary point* to be a first-order stationary point $\mathbf{x}$ that additionally satisfies $\nabla^2 f(\mathbf{x}) \succ 0$, which is a sufficient condition for $\mathbf{x}$ to be a local minimizer. Our main contribution is showing that under suitable assumptions, a perturbed version of proximal gradient descent will generate a sequence that converges to an approximate second-order stationary point. We make assumptions on the second-order behavior of f, similar to assumptions under which it is known that gradient descent will always converge to a second-order stationary point except for adversarially chosen starting points [14] — in contrast to Newton's method, which is attracted to all stationary points. However, even in the smooth case when the sequence converges, gradient descent converges arbitrarily slowly [10] in the presence of a saddle point, so perturbation is necessary. In the non-smooth case, perturbation is even more important due to the proximal nature of the algorithm.

A toy example: Gaussian Bump Consider the function $\Phi : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{1}{2}(x^2 - y^2)\mathrm{e}^{-\frac{x^2+y^2}{5}} + \frac{1}{100}h_{100}(\mathbf{x})$ where $h_{100}(\mathbf{x})$ is the Huber function with parameter 100 [3]. The choice of this combination of Huber parameter and the magnitude of Huber function ensures that the origin is a saddle point. The Huber function approximates the ℓ_1 norm. The plot is show in Fig. 1.

This function has two local minima and a saddle point at $(0, 0)$. Because the Huber function is both smooth and it has a known proximity operator, we can treat it as either part of the smooth f component or the non-smooth g component, and therefore run either gradient descent or proximal gradient descent. We experiment with both algorithms, randomly picking initial points at $\mathbf{x}_0 = (0.3, 0.01) + \boldsymbol{\xi}$ where $\boldsymbol{\xi}$ is sampled uniformly from $\mathbb{B}_0(\frac{1}{10}\|\mathbf{x}_0\|)$, and varying the stepsize η, with fixed maximum iteration 1000. Figure 2 shows the empirical success rate of finding a local minimizer (as opposed to converging to the saddle point at $(0, 0)$).

We observe that the range of stable step size for the proximal descent algorithm is wider than gradient descent, and the success rate of proximal descent

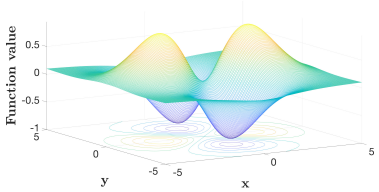


Fig. 1: Graph of function $\Phi(\mathbf{x})$

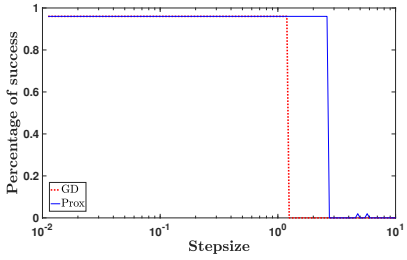


Fig. 2: The comparison between gradient descent (GD) and proximal gradient descent (Prox) on the percentage of success finding the correct local minima, as a function of the stepsize η

is as high as the gradient descent. This example motivates us to adopt proximal descent over gradient descent in real application for better stability and equivalent, if not better, accuracy.

A coincidence In this toy example, the saddle point at $(0, 0)$ happens to be a fixed point of proximal operator of $\eta\lambda\|\mathbf{x}\|_1$. Soft thresholding, as the proximal operator of $\lambda\|\mathbf{x}\|_1$ is known [7], has an attracting region that sets nearby points to 0. The radius of the attracting region (per dimension) is $\eta\lambda$, thus if $\|\mathbf{x}_{t_0} - \eta\nabla f(\mathbf{x}_{t_0})\|_\infty \leq \eta\lambda$ for some iteration t_0, then $\mathbf{x}_t = 0$ for all $t > t_0$. Proximal gradient descent performs even better when the saddle point is not in the attracting region.

Structure of the paper Section 2 states the algorithm, followed by section 3 where the theoretical guarantee is presented with proof. Section 4 shows numerical experiments.

1.1 Related literature

Second order methods for smooth objectives Some recent second order methods, mainly based on either cubic-regularized Newton methods as in [17] or based on trust-region methods (as in Curtis et al. [8]), have been shown to converge to ε-approximate local minimizers of smooth non-convex objective functions in $\mathcal{O}(\varepsilon^{-1.5})$ iterations. See [6, 13, 21] for a more thorough review of these methods. We do not consider these methods further due to the high-cost of solving for the Newton step in large dimensions.

First order methods for smooth objectives We focus on first order methods because each step is cheaper and these methods are more frequently adopted by the deep learning community. Xu et al. in [20] and Allen-Zhu et al. in [21] develop Negative-Curvature (NC) search algorithms, which find descent direction corresponding to negative eigenvalues of Hessian matrix. The NC search routines

avoid using either Hessian or Hessian-vector information directly, and it can be applied in both online and deterministic scenarios. In the online setting, combining NC search routine with first-order stochastic methods will give algorithms NEON-$\mathcal{A}$ [20] and NEON2+SGD [21] with iteration cost $\mathcal{O}(\frac{d}{\varepsilon^{3.5}})$ and $\mathcal{O}(\varepsilon^{-3.5})$ respectively (the latter still depends on dimension, whose induced complexity is at least $\ln^2(d)$), and these methods generate a sequence that converges to an approximate local minimum with high probability. In the offline setting, Jin et al. in [13] provide a stochastic first order method that finds an approximate local minimizer with high probability at computational cost $\mathcal{O}(\frac{\ln^4(d)}{\varepsilon^2})$. Combining NEON2 with gradient descent or SVRG, the cost to find an approximate local minimum is $\mathcal{O}(\varepsilon^{-2})$, whose dependence on dimension is not specified but at least $\ln^2(d)$. These methods make Lipschitz continuity assumptions about the gradient and Hessian, so they do not apply to non-smooth optimization.

A recent preprint [15] approaches the problem of finding local minima using the forward-backward envelope technique developed in [19], where the assumption about the smoothness of objective function is weakened to local smoothness instead of global smoothness.

Non-smooth objectives In the offline settings, Boţ et al. propose a proximal algorithm for minimizing non-convex and non-smooth objective functions in [5]. They show the convergence to KKT points instead of approximate second-order stationary points. Other work [1, 4] relies on the Kurdya-Lojasiewicz inequality and shows convergence to stationary points in the sense of the limiting subdifferential, which is not the same as a local minimizer or approximate second-order stationary point. In the online setting, Reddi et al. demonstrated in [12] that the proximal descent with variance reduction technique (proxSVRG) has linear convergence to a first-order stationary point, but not to a local minimizer.

2 Algorithm

The algorithm takes as input a starting vector $\mathbf{x}_0$, the gradient Lipschitz constant L, the Hessian Lipschitz constant ρ, the second-order stationary point tolerance ε, a positive constant c, a failure probability δ, and estimated function value gap Δ_Φ. The key parameter for Algorithm 1 is the constant c. It should be made large enough so that the effect of perturbation will be significant enough for escaping saddle points, and at the same time not too large so that the iteration stepsize is of reasonable magnitude and the iteration will not go wild. The output of the algorithm is an ε-second-order stationary point (see Def. 3).

Algorithm 1 Perturbed Proximal Descent: input$(\mathbf{x}_0, L, \rho, \varepsilon, c, \delta, \Delta_\Phi)$

$\chi \leftarrow 3\max\{\ln(\frac{dL\Delta_\Phi}{c\varepsilon^2\delta}), 4\},\ \eta \leftarrow \frac{c}{L},\ r \leftarrow \frac{\sqrt{c}}{\chi^2}\cdot\frac{\varepsilon}{L},\ g_{\text{thres}} \leftarrow \frac{\sqrt{c}}{\chi^2}\cdot\varepsilon,\ \Phi_{\text{thres}} \leftarrow \frac{c}{\chi^3}\cdot\sqrt{\frac{\varepsilon^3}{\rho}},\ t_{\text{thres}} \leftarrow \frac{\chi}{c^2}\cdot\frac{L}{\sqrt{\rho\varepsilon}}$
$t_{\text{noise}} \leftarrow -t_{\text{thres}} - 1$
for $t = 0, 1, \ldots$ **do**
  if $\|\mathbf{x} - \text{prox}_{\eta g}[\mathbf{x} - \eta\nabla f(\mathbf{x})]\| < g_{\text{thres}}$ and $t - t_{\text{noise}} > t_{\text{thres}}$ **then**
    $\tilde{\mathbf{x}}_t \leftarrow \mathbf{x}_t, \quad t_{\text{noise}} \leftarrow t$
    $\mathbf{x}_t \leftarrow \tilde{\mathbf{x}}_t + \boldsymbol{\xi}_t, \qquad \boldsymbol{\xi}_t$ uniformly $\sim \mathbb{B}_0(r)$
  if $t - t_{\text{noise}} = t_{\text{thres}}$ and $\Phi(\mathbf{x}_t) - \Phi(\tilde{\mathbf{x}}_{t_{\text{noise}}}) > -\Phi_{\text{thres}}$ **then**
    return $\tilde{\mathbf{x}}_{t_{\text{noise}}}$
  $\mathbf{x}_{t+1} \leftarrow \text{prox}_{\eta g}[\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)]$

3 Escaping Saddle Points through Perturbed Proximal Descent

The main step in the algorithm is a proximal gradient descent step applied to $f + g$, defined as

$$\begin{aligned}\mathbf{x}_{t+1} &= \operatorname*{argmin}_{\mathbf{y}} f(\mathbf{x}_t) + \langle\nabla f(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t\rangle + \frac{\eta^{-1}}{2}\|\mathbf{y} - \mathbf{x}_t\|^2 + g(\mathbf{y}) \\ &= \text{prox}_{\eta g} \circ (I - \eta\nabla f)(\mathbf{x}_t) \end{aligned} \tag{2}$$

One motivation of preferring proximal descent to gradient descent, as shown in Figure 2, is the stability of the algorithm with respect to stepsize change. The proximal step is similar to the implicit/backward Euler scheme, as equation (2) can be written as $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\big(\nabla f(\mathbf{x}_t) + \partial g(\mathbf{x}_{t+1})\big)$. From this perspective, we expect that proximal descent will demonstrate at least the same convergence speed as gradient descent and stronger stability with respect to hyperparameter setting.

Definition 1 (Gradient Mapping). *Consider a function $\Phi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$. The gradient mapping is defined as $G_\eta^{f,g}(\mathbf{x}) := \mathbf{x} - \text{prox}_{\eta g}[\mathbf{x} - \eta\nabla f(\mathbf{x})]$*

In the rest of this paper, the super- and subscript of the gradient mapping are not specified, as it is always clear that f represents the smooth nonconvex part of Φ, g represents $\lambda\|\mathbf{x}\|_1$, and η is the stepsize used in the algorithm. Observe that the gradient map is just the gradient of f if $g \equiv 0$.

Definition 2 (First order stationary points). *For a function $\Phi(\mathbf{x})$, define first order stationary points as the points which satisfy $G(\mathbf{x}) = 0$.*

Definition 3 (ε-second-order stationary point). *Consider a function $\Phi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$. A point $\mathbf{x}$ is an ε-second-order stationary point if*

$$\|G(\mathbf{x})\| \le \varepsilon \quad \textit{and} \quad \lambda\big(\nabla^2 f(\mathbf{x})\big)_{\min} \ge -\sqrt{\rho\varepsilon} \tag{3}$$

where $\lambda(\cdot)_{\min}$ is the smallest eigenvalue.

The first Lipschitz assumption below is standard [3], and the assumption on the Hessian was used in [13] (for example, it is true if f is quadratic).

Assumption A1 (Lipschitz Properties) *∇f is L-Lipschitz continuous and $\nabla^2 f$ is ρ Lipschitz continuous. We write $\mathcal{H}$ as shorthand for $\nabla^2 f(\mathbf{x})$ when $\mathbf{x}$ is clear from context.*

Assumption A2 (Moderate Nonsmooth Term) *The magnitude of $\|\mathbf{x}\|_1$ term, which is denoted by λ, satisfies inequalities (7) and (9).*

Theorem 1 (Main). *There exists an absolute constant $c_{\max}$ such that if $f(\cdot)$ satisfies A1 and A2, then for any $\delta > 0, \varepsilon \leq \frac{L^2}{\rho}, \Delta_\Phi \geq \Phi(\mathbf{x}_0) - \Phi^\star$, and constant $c \leq c_{\max}$, with probability $1-\delta$, the output of PPD$(\mathbf{x}_0, L, \rho, \varepsilon, c, \delta, \Delta_f)$ will be a ε-second order stationary point, and terminate in iterations:*

$$\mathcal{O}\left(\frac{L(\Phi(\mathbf{x}_0) - \Phi^\star)}{\varepsilon^2} \ln^4\left(\frac{dL\Delta_\Phi}{\varepsilon^2\delta}\right)\right)$$

Remark Assuming $\varepsilon \leq \frac{L^2}{\rho}$ does not lead to loss of generality. Recall the second order condition is specified as $\lambda\big(\nabla^2 f(\mathbf{x}^\star)\big)_{\min} \geq -\sqrt{\rho\varepsilon}$, since when $\varepsilon \geq \frac{L^2}{\rho}$, we always have $-\sqrt{\rho\varepsilon} \leq -L \leq \lambda\big(\nabla^2 f(\mathbf{x}^\star)\big)_{\min}$, where the second inequality follows from the fact that the Lipschitz constant is the upper bound for $\lambda(\nabla^2 f(\mathbf{x}))$ in norm. Consequently, when $\varepsilon \geq \frac{L^2}{\rho}$, every ε-second-order stationary point is automatically a first order stationary point.

For the proof of the main theorem, we introduce some notation and units for the simplicity of proof statement.

For matrices we use $\|\cdot\|$ to denote spectral norm. The operator $\mathcal{P}_{\mathcal{S}}(\cdot)$ denotes projection onto set $\mathcal{S}$. Define the local approximation of the smooth part of the objective function by

$$\tilde{f}_{\mathbf{x}}(\mathbf{y}) := f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{z})^T\mathcal{H}(\mathbf{y}-\mathbf{z}) \tag{4}$$

Units With $\kappa := \frac{L}{\gamma} \geq 1$, we define the following units for the convenience of proof statement:

$$\mathscr{F} := \eta L\frac{\gamma^3}{\rho^2} \cdot \ln^{-3}(\frac{d\kappa}{\delta}), \qquad \mathscr{G} := \sqrt{\eta L}\frac{\gamma^2}{\rho} \cdot \ln^{-2}(\frac{d\kappa}{\delta})$$

$$\mathscr{S} := \sqrt{\eta L}\frac{\gamma}{\rho} \cdot \ln^{-1}(\frac{d\kappa}{\delta}), \qquad \mathscr{T} := \frac{\ln(\frac{d\kappa}{\delta})}{\eta\gamma}$$

3.1 Lemma: Iterates remain bounded if stuck near a saddle point

Lemma 1. *For any constant $\hat{c} \geq 3$, there exists absolute constant $c_{\max}$: for any $\delta \in (0, \frac{d\kappa}{e}]$, let $f(\cdot), \tilde{\mathbf{x}}$ satisfies the condition in Lemma 6, for any initial point $\mathbf{u}_0$ with $\|\mathbf{u}_0 - \tilde{\mathbf{x}}\| \leq 2\mathscr{S}/(\kappa \cdot \ln(\frac{d\kappa}{\delta}))$, define:*

$$T = \min\left\{ \inf_t \left\{ t \,|\, \tilde{f}_{\mathbf{u}_0}(\mathbf{u}_t) - f(\mathbf{u}_0) + g(\mathbf{u}_t) - g(\mathbf{u}_0) \leq -3\mathscr{F} \right\}, \hat{c}\mathscr{T} \right\}$$

then, for any $\eta \le c_{\max}/L$, we have for all $t < T$ that $\|\mathbf{u}_t - \tilde{\mathbf{x}}\| \le 100(\mathscr{S} \cdot \hat{c})$.

Proof. We show if the function value did not decrease, then all the iteration updates must be constrained in a small ball. The proximal descent updates the solution as

$$\begin{aligned}\tilde{\mathbf{u}}_{t+1} &= \mathbf{u}_t - \nabla f(\mathbf{u}_t) = (I - \nabla f)(\mathbf{u}_t)\\ \mathbf{u}_{t+1} &= \mathrm{prox}_{\eta g}\big(\tilde{\mathbf{u}}_{t+1}\big) = \mathrm{prox}_{\eta g} \circ (I - \nabla f)(\mathbf{u}_t)\end{aligned}$$

Without losing of generality, set $\mathbf{u}_0 = 0$ to be the origin. For any $t \in \mathbb{N}$,

$$\|\mathbf{u}_t - \mathbf{u}_0\| = \|\mathbf{u}_t - 0\| = \|\mathrm{prox}_{\eta g}(\tilde{\mathbf{u}}_t) - \mathrm{prox}_{\eta g}(0)\| \le \|\tilde{\mathbf{u}}_t - 0\| = \|\tilde{\mathbf{u}}_t\|$$

Jin et al. prove in [13] by induction that if $\|\mathbf{u}_t\| \le 100(\mathscr{S} \cdot \hat{c})$, then $\|\tilde{\mathbf{u}}_{t+1}\| \le 100(\mathscr{S} \cdot \hat{c})$. Consequently, $\|\mathbf{u}_{t+1}\| \le 100(\mathscr{S} \cdot \hat{c})$.

We point out that it is implicitly assumed that $\frac{2\mathscr{S}}{\kappa \cdot \ln(\frac{d\kappa}{\delta})} \ll \hat{c}$, so that for all $t < T$, $\|\tilde{\mathbf{x}}\| \ll \|\mathbf{u}_t\|$, and the relation $\|\mathbf{u}_t - \tilde{\mathbf{x}}\| \le \|\mathbf{u}_t\| + \|\tilde{\mathbf{x}}\| \le 100(\mathscr{S} \cdot \hat{c})$ holds.

3.2 Preparation for Building Pillars

Lemma 2 (Existence of lower bound for the difference sequence $\{\mathbf{v}_t\}_{t=1}^T$). *For iteration sequences $\{\mathbf{w}_t\}$ and $\{\mathbf{u}_t\}$ defined in Lemma 4, define the difference sequence as*

$$\mathbf{v}_t = \mathbf{w}_t - \mathbf{u}_t$$

There exists a positive lower bound for $\{\mathbf{v}_t\}$ when $t < \hat{c}\mathscr{T}$.

Proof. To show that the lower bound for iteration difference $\{\mathbf{v}_t\}_{t=1}^T$ exists, we consider bounding the iteration sequence $\tilde{\mathbf{v}}_{t+1}$ first. Define the difference between the proximal of l_1 penalty term and its coimage as $\mathcal{D}_g[\mathbf{x}] = \mathrm{prox}_g[\mathbf{x}] - \mathbf{x} = \min\{\lambda \mathbb{1}, |\mathbf{x}|\} \otimes \mathrm{sgn}(-\mathbf{x})$, where $\otimes$ is Hadamard product and the minimum is taken elementwise. We notice that $\|\mathcal{D}_{\eta\lambda\|\cdot\|_1}[\mathbf{x}]\| \le \eta\lambda\sqrt{d}$. Thus, $\|\mathbf{w}_k - \mathbf{u}_k\| = \|\tilde{\mathbf{w}}_k - \tilde{\mathbf{v}}_k - \lambda(\mathcal{D}_{\eta g}[\tilde{\mathbf{w}}_k] - \mathcal{D}_{\eta g}[\tilde{\mathbf{u}}_k])\| \ge \|\tilde{\mathbf{w}}_k - \tilde{\mathbf{v}}_k\| - 2\eta\lambda\sqrt{d}$.

$$\begin{aligned}\|\tilde{\mathbf{v}}_{t+1}\| &= \|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{u}}_{t+1}\|\\ &= \|(I - \eta\nabla f) \circ \mathrm{prox}_{\eta g}(\tilde{\mathbf{w}}_k) - (I - \eta\nabla f) \circ \mathrm{prox}_{\eta g}(\tilde{\mathbf{u}}_k)\|\\ &= \|\mathbf{w}_k - \mathbf{u}_k - \eta(\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{u}_k))\|\\ &\ge \|\mathbf{w}_k - \mathbf{u}_k\| - \eta L\|\mathbf{w}_k - \mathbf{u}_k\| = (1 - \eta L)\|\mathbf{w}_k - \mathbf{u}_k\|\\ &\ge (1 - \eta L)(\|\tilde{\mathbf{w}}_k - \tilde{\mathbf{u}}_k\| - 2\eta\lambda\sqrt{d}) = (1 - \eta L)(\|\tilde{\mathbf{v}}_k\| - 2\eta\lambda\sqrt{d})\\ &\ge (1 - \eta L)^t\|\tilde{\mathbf{v}}_1\| - 2\eta\lambda\sqrt{d}\sum_{i=1}^{t}(1 - \eta L)^i\\ &= (1 - \eta L)^t\|\tilde{\mathbf{v}}_1\| - 2\lambda\sqrt{d}\frac{(1 - \eta L)\big(1 - (1 - \eta L)^t\big)}{L}\end{aligned}$$

As $\tilde{\mathbf{v}}_1 = (I-\eta\nabla f)\mathbf{v}_0 = (I-\eta\nabla f)\mu r\mathbf{e}_1 = \mu r(\mathbf{e}_1 - \eta\nabla^2 f(\boldsymbol{\xi})\theta\mathbf{e}_1) = \mu r(1+\eta\gamma\theta)\mathbf{e}_1$, where $\theta \in (0,1)$, we have

$$\|\tilde{\mathbf{v}}_{t+1}\| \geq (1-\eta L)^t \mu r(1+\eta\gamma\theta) - 2\lambda\sqrt{d}\frac{(1-\eta L)(1-(1-\eta L)^t)}{\eta L} \tag{5}$$

To compare $\|\mathbf{v}_t\|$ and $\|\tilde{\mathbf{v}}_t\|$,

$$\|\mathbf{v}_{t+1}\| \geq \|\tilde{\mathbf{v}}_{t+1}\| - 2\eta\lambda\sqrt{d} \geq (1-\eta L)^t \mu r(1+\eta\gamma\theta) - 2\lambda\sqrt{d}\frac{(1-\eta L)(1-(1-\eta L)^t)+\eta L}{L} \tag{6}$$

Therefore, as long as

$$\lambda < \frac{(1-\eta L)^{\hat{c}_{\mathscr{T}}}\mu\frac{1}{\kappa(\ln\frac{d\kappa}{\delta})^2}\sqrt{\eta}L^{\frac{3}{2}}\frac{\gamma}{\rho}(1+\eta\gamma\theta)}{2\sqrt{d}[(1-\eta L)(1-(1-\eta L)^{\hat{c}_{\mathscr{T}}})+\eta L]} \tag{7}$$

the difference sequence $\{\|\mathbf{v}_t\|\}$ has a positive lower bound on its norm.

Lemma 3 (Preservation of subspace projection monotonicity after prox of l_1 in rotated coordinate with small λ).

Denote the subspace of $\mathbb{R}^n$ spanned by $\{\mathbf{e}_1\}$ as $\mathbb{E}$, while the complement subspace spanned by $\{\mathbf{e}_2, \cdots, \mathbf{e}_n\}$ as $\mathbb{E}^\perp$. For a given vector $\mathbf{x}$ chosen from a lower bounded set $\mathcal{X}$, i.e. $\forall\,\mathbf{x}\in\mathcal{X}$, $\|\mathbf{x}\| \geq C$ for some constant $C>0$, assume $\|\mathcal{P}_{\mathbb{E}^\perp}\mathbf{x}\| \leq K\|\mathcal{P}_{\mathbb{E}}\mathbf{x}\|$, where $0<K\leq 1$ is a constant. If the parameter λ for the l_1 penalty term is small enough, then

$$\|\mathcal{P}_{\mathbb{E}^\perp}\mathrm{prox}_{\eta g}(\mathbf{x})\| \leq K\|\mathcal{P}_{\mathbb{E}}\mathrm{prox}_{\eta g}(\mathbf{x})\|$$

Proof. We want to find a constraint on λ such that when λ is small enough, if the projection in the original coordinate demonstrates the monotonicity relation $\|\mathcal{P}_{\mathbb{E}}\mathbf{x}\| \leq \|\mathcal{P}_{\mathbb{E}^\perp}\mathbf{x}\|$, this monotonicity relation will be preserved after proximal operator of l_1 is applied on the input vector.

Naturally there exists a normal vector, denoted as $\hat{\boldsymbol{n}}_{\text{boundary}} \equiv \hat{\boldsymbol{n}}$, for the boundary hyperplane on which $\|\mathcal{P}_{\mathbb{E}}\mathbf{x}\| = K\|\mathcal{P}_{\mathbb{E}^\perp}\mathbf{x}\|$. By moving along $\hat{\boldsymbol{n}}$, a point approaches the boundary most efficiently. Any vector inside the hyperplane is perpendicular to $\hat{\boldsymbol{n}}$, which we denote as $\hat{\boldsymbol{n}}^\perp$.

Define

$$\hat{\mathbf{v}}_{\text{move}}(\mathbf{x}) = \begin{cases} -\eta\lambda\cdot\mathrm{sgn}(x_i) & \text{if } |x_i| \geq \eta\lambda \\ -x_i & \text{if } |x_i| < \eta\lambda \end{cases} = \min\{|x|, \eta\lambda\mathbb{1}\} \otimes \mathrm{sgn}(-\mathbf{x}) \tag{8}$$

where $\otimes$ is the Hadamard product, and the minimum is taken elementwise. Because $\mathrm{prox}_{\eta g}(\mathbf{x}) = \mathbf{x} + \hat{\mathbf{v}}_{\text{move}}$, a sufficient condition to be imposed on λ to guarantee the preservation of projection monotonicity $\|\mathcal{P}_{\mathbb{E}^\perp}\mathrm{prox}_{\eta g}(\mathbf{x})\| \leq K\|\mathcal{P}_{\mathbb{E}}\mathrm{prox}_{\eta g}(\mathbf{x})\|$ is that

$$\lambda < \left\|\frac{\mathrm{Proj}_{\boldsymbol{n}}\mathbf{x}}{\hat{\mathbf{v}}_{\text{move}}\cdot\hat{\boldsymbol{n}}}\right\| = \left\|\frac{\mathbf{x}\cdot\hat{\boldsymbol{n}}}{\hat{\mathbf{v}}_{\text{move}}\cdot\hat{\boldsymbol{n}}}\right\| \leq \frac{\|\mathbf{x}\|}{\|\hat{\mathbf{v}}_{\text{move}}\cdot\hat{\boldsymbol{n}}\|}$$

which means the moving distance caused by applying the l_1 proximal operator (soft shrinkage) projected on the direction of $\hat{\boldsymbol{n}}$ is less that the distance between $\mathbf{x}$ to the boundary hyperplane, hence rendering the vector stay on the same side of the boundary after moving.

Therefore, as long as

$$\lambda < \frac{C}{\|\hat{\mathbf{v}}_{\text{move}} \cdot \hat{\boldsymbol{n}}\|} \tag{9}$$

the monotonicity of projection onto subspaces can be preserved.

Remark 1 for Lemma 3 As an examples in $\mathbb{R}^2$, set $K = 1$, we visualise the shift caused by proximal operator and the boundary of projection-monotonicity preserving region. Assume $\mathbf{e}_{1,2}$ are orthonormal basis of Cartesian coordinate in the standard position. The directional vector for region division boundary is $\hat{\mathbf{e}}_{\text{boundary}} = \hat{\boldsymbol{n}}^{\perp} = \frac{\pm\hat{\mathbf{e}}_1 \pm \hat{\mathbf{e}}_2}{\sqrt{2}}$, and $\hat{\mathbf{e}}^{\perp}_{\text{boundary}} = \hat{\boldsymbol{n}}$ is the corresponding perpendicular directional vector. For l_1 norm, $\hat{\mathbf{v}}_{\text{move}}$ is $(\pm 1, \pm 1)$.

Remark 2 for Lemma 3 We point out that the upper bound for the parameter λ is related to the alignment of the eigenspace of $\mathcal{H}$. If the eigenspace of $\mathcal{H}$ is aligned with canonical orthonormal basis of $\mathbb{R}^d$, then $\lambda \in (0, \infty)$. The most stringent restriction on the upper bound of λ applies when $\hat{\mathbf{v}}_{\text{move}}$ is parallel to $\hat{\boldsymbol{n}}$.

3.3 Lemma: Perturbed iterates will escape the saddle point

Lemma 4. *There exists absolute constant $c_{\max}, \hat{c}$ such that: for any $\delta \in (0, \frac{d\kappa}{e}]$, let $f(\cdot), \tilde{\mathbf{x}}$ satisfies the condition in Lemma 6, and sequences $\{\mathbf{u}_t\}, \{\mathbf{w}_t\}$ satisfy the conditions in Lemma 6, define:*

$$T = \min\left\{ \inf_t \left\{ t | \tilde{f}_{\mathbf{w}_0}(\mathbf{w}_t) + g(\mathbf{w}_t) - f(\mathbf{w}_0) - g(\mathbf{w}_0) \leq -3\mathscr{F} \right\}, \hat{c}\mathscr{T} \right\}$$

then, for any $\eta \leq c_{\max}/L$, if $\|\mathbf{u}_t - \tilde{\mathbf{x}}\| \leq 100(\mathscr{S} \cdot \hat{c})$ for all $t < T$, we will have $T < \hat{c}\mathscr{T}$.

Proof. We show that if the iterate sequence before time T starting from $\mathbf{u}_0$ does not provide sufficient function value decrease, the other iterate sequence, which starts from $\mathbf{w}_0$, will be able to achieve the function value decrease purpose. Ultimately, we will prove $T < \hat{c}\mathscr{T}$. We establish the inequality about T by considering the difference between $\mathbf{w}_t$ and $\mathbf{u}_t$. Define $\mathbf{v}_t = \mathbf{w}_t - \mathbf{u}_t$. The assumption of the lemma 4, $\mathbf{v}_0 = \mu[\mathscr{S}/(\kappa \cdot \ln(\frac{d\kappa}{\delta}))]\mathbf{e}_1$, $\mu \in [\delta/(2\sqrt{d}), 1]$.

We bound $\|\mathbf{v}_t\|$ from both sides for all $t < T$ to obtain an inequality about T.

Recall that the proximal descent updates the solution as

$$\tilde{\mathbf{u}}_{t+1} = \mathbf{u}_t - \nabla f(\mathbf{u}_t) = (I - \eta\nabla f)(\mathbf{u}_t)$$
$$\mathbf{u}_{t+1} = \mathrm{prox}_{\eta g}\big(\tilde{\mathbf{u}}_{t+1}\big) = \mathrm{prox}_{\eta g} \circ (I - \eta\nabla f)(\mathbf{u}_t)$$

Simple algebraic computation gives

$$\tilde{\mathbf{v}}_{t+1} = (I - \eta\mathcal{H} - \eta\Delta'_t)\mathbf{v}_t \tag{10}$$

where $\Delta'_t = \int_0^1 \nabla^2 f(\mathbf{u}_t + \theta\mathbf{v}_t)\,\mathrm{d}\theta - \mathcal{H}$, and $\tilde{\mathbf{v}}_t = \tilde{\mathbf{w}}_t - \tilde{\mathbf{u}}_t$.

Consider $\|\tilde{\mathbf{u}}_t\|$ and $\|\tilde{\mathbf{w}}_t\|$. Because $\mathbf{v}_0 = \tilde{\mathbf{v}}_0$, we have $\|\tilde{\mathbf{w}}_0 - \tilde{\mathbf{x}}\| \le \|\tilde{\mathbf{u}}_0 - \tilde{\mathbf{x}}\| + \|\tilde{\mathbf{v}}_0\| \le 2\mathscr{S}/(\kappa \cdot \ln(\frac{d\kappa}{\delta}))$. With same logic in the proof for lemma 1, we see $\|\tilde{\mathbf{u}}_t\| \le 100(\mathscr{S} \cdot \hat{c})$, and $\|\tilde{\mathbf{w}}_t\| \le 100(\mathscr{S} \cdot \hat{c})$. (Same relation hold for $\|\mathbf{u}_t\|$ and $\|\mathbf{w}_t\|$ respectively.) As a result, $\|\tilde{\mathbf{v}}_t\| \le \|\tilde{\mathbf{w}}_t\| + \|\tilde{\mathbf{u}}_t\| \le 200(\mathscr{S} \cdot \hat{c})$ for all $t < T$. Also,

$$\|\mathbf{v}_t\| \le 200(\mathscr{S} \cdot \hat{c}) \tag{11}$$

Equation (11) and Hessian Lipschitz gives for $t < T$, $\|\Delta'_t\| \le \rho(\|\mathbf{u}_t\| + \|\mathbf{v}_t\| + \|\tilde{\mathbf{x}}\|) \le \rho\mathscr{S}(300\hat{c} + 1) = \frac{\zeta}{\eta}$, where $\zeta = \eta\rho\mathscr{S}(300\hat{c} + 1)$.

Denote ψ_t be the norm of $\mathbf{v}_t$ projected onto $\mathbf{e}_1$ direction ($\mathcal{S}$), and φ_t be the norm of $\mathbf{v}_t$ projected onto the remaining subspace ($\mathcal{S}^c$), while $\tilde{\psi}_t$ be the norm of $\tilde{\mathbf{v}}_t$ projected onto $\mathcal{S}$, and $\tilde{\varphi}_t$ be the norm of $\tilde{\mathbf{v}}_t$ projected onto $\mathcal{S}^c$.

Equation (10) gives

$$\tilde{\psi}_{t+1} \ge (1 + \gamma\eta)\psi_t - \zeta\sqrt{\psi_t^2 + \varphi_t^2} \tag{12}$$

$$\tilde{\varphi}_{t+1} \le (1 + \gamma\eta)\varphi_t + \zeta\sqrt{\psi_t^2 + \varphi_t^2} \tag{13}$$

To obtain the lower bound of $\|\mathbf{v}_t\|$, we prove the following relation as preparation:

$$\text{for all } t < T, \quad \varphi_t \le 4\zeta t \cdot \psi_t \tag{14}$$

By hypothesis of lemma 4, we know $\varphi_0 = 0$, thus the base case of induction holds. Assume equation (14) is true for $\tau \le t$, for $t + 1 \le T$, we have

$$\tilde{\varphi}_{t+1} \le 4\zeta t(1 + \gamma\eta)\psi_t + \zeta\sqrt{\psi_t^2 + \varphi_t^2}$$
$$4\zeta(t+1)\left[(1 + \gamma\eta)\psi_t - \zeta\sqrt{\psi_t^2 + \varphi_t^2}\right] \le 4\zeta(t+1)\tilde{\psi}_{t+1} \tag{15}$$

By choosing $\sqrt{c_{\max}} \le \frac{1}{300\hat{c}+1}\min\{\frac{1}{2\sqrt{2}}, \frac{1}{4\hat{c}}\}$, and $\eta \le \frac{c_{\max}}{L}$, we have $4\zeta(t+1) \le 4\zeta T \le 4\eta\rho\mathscr{S}(300\hat{c} + 1)\hat{c}\mathscr{T} = 4\sqrt{\eta L}(300\hat{c} + 1)\hat{c} \le 1$. This gives $4(1 + \gamma\eta)\psi_t \ge 4\psi_t \ge (1 + 1)\sqrt{2\psi_t^2} \ge (1 + 4\zeta(t + 1))\sqrt{\psi_t^2 + \varphi_t^2}$. i.e.

$$(1 + 4\zeta(t + 1))\sqrt{\psi_t^2 + \varphi_t^2} \le 4\psi_t \tag{16}$$

Connecting two parts of equation (15), we obtain

$$\tilde{\varphi}_{t+1} \leq 4\zeta(t+1)\tilde{\psi}_{t+1} \tag{17}$$

Now we switch our focus to the eigenspace of Hessian $\mathcal{H}$. Assume the orthonormal basis for the eigensapce of $\mathcal{H}$ is $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_d\}$. The order of dimension aligns with the increasing order of the corresponding eigenvalues. This coordinate transformation does not lead to loss of generality, as it is unitary. By lemma 2, we know the iteration difference sequence $\mathbf{v}_t$ has a positive lower bound in terms of 2-norm. Therefore, by lemma 3, with the virtue of equation (17) $\sqrt{\sum_{i=2}^{d}(\mathbf{e}_i^T\tilde{\mathbf{v}}_{t+1})^2} \leq 4\zeta(t+1)\|\mathbf{e}_1^T\tilde{\mathbf{v}}_{t+1}\|$, we still have the projection monotonicity on the subspace of eigenspace of $\mathcal{H}$, i.e.

$$\varphi_{t+1} = \sqrt{\sum_{i=2}^{d}(\mathbf{e}_i^T\operatorname{prox}_g(\tilde{\mathbf{v}}_{t+1}))^2} \leq 4\zeta(t+1)\|\mathbf{e}_1^T\operatorname{prox}_g(\tilde{\mathbf{v}}_{t+1})\| = 4\zeta(t+1)\psi_{t+1}$$

Until here we finish the induction.

Recall that $4\zeta(t+1) \leq 1$, we thus have $\varphi_t \leq 4\zeta t\psi_t \leq \psi_t$, which gives

$$\psi_{t+1} \geq (1+\gamma\eta)\psi_t - \sqrt{2}\zeta\psi_t \geq \big(1+\frac{\gamma\eta}{2}\big)\psi_t \tag{18}$$

where the last inequality follows from $\zeta = \eta\rho\mathscr{S}(300\hat{c}+1) \leq \sqrt{c_{\max}}(300\hat{c}+1)\gamma\eta \cdot \ln^{-1}(\frac{d\kappa}{\delta}) \leq \frac{\gamma\eta}{2\sqrt{2}}$.

Finally, combining (11) and (18), we have for all $t < T$:

$$\begin{aligned} 200(\mathscr{S}\cdot\hat{c}) \geq \|\mathbf{v}_t\| \geq \psi_t &\geq (1+\frac{\gamma\eta}{2})^t\psi_0 = (1+\frac{\gamma\eta}{2})^t c_0\frac{\mathscr{S}}{\kappa}\ln^{-1}\left(\frac{d\kappa}{\delta}\right) \\ &\geq (1+\frac{\gamma\eta}{2})^t\frac{\delta}{2\sqrt{d}}\frac{\mathscr{S}}{\kappa}\ln^{-1}\left(\frac{d\kappa}{\delta}\right) \end{aligned}$$

This implies

$$T < \frac{1}{2}\frac{\ln[400\frac{\kappa\sqrt{d}}{\delta}\cdot\hat{c}\ln(\frac{d\kappa}{\delta})]}{\ln(1+\frac{\gamma\eta}{2})} \leq \frac{\ln[400\frac{\kappa\sqrt{d}}{\delta}\cdot\hat{c}\ln(\frac{d\kappa}{\delta})]}{\gamma\eta} \leq (2+\ln(400\hat{c}))\mathscr{T}$$

The last inequality is due to $\delta \in (0, \frac{d\kappa}{e}]$, we have $\ln(\frac{d\kappa}{\delta}) \geq 1$. By choosing the constant $\hat{c}$ to be large enough to satisfy $2+\ln(400\hat{c}) \leq \hat{c}$, we will have $T < \hat{c}\mathscr{T}$, which finishes the proof.

3.4 Combining Previous Results

Lemma 5. *There exists a universal constant $c_{\max}$, for any $\delta \in (0, \frac{d\kappa}{e}]$, let $f(\cdot), \tilde{\mathbf{x}}$ satisfies the conditions in Lemma 6, and without loss of generality let*

$\mathbf{e}_1$ *be the minimum eigenvector of* $\nabla^2 f(\tilde{\mathbf{x}})$*. Consider two gradient descent sequences* $\{\mathbf{u}_t\}, \{\mathbf{w}_t\}$ *with initial points* $\mathbf{u}_0, \mathbf{w}_0$ *satisfying: (denote radius* $r = \mathscr{S}/(\kappa \cdot \ln(\frac{d\kappa}{\delta}))$*)*

$$\|\mathbf{u}_0 - \tilde{\mathbf{x}}\| \le r, \quad \mathbf{w}_0 = \mathbf{u}_0 + \mu \cdot r \cdot \mathbf{e}_1, \quad \mu \in [\delta/(2\sqrt{d}), 1]$$

Then, for any stepsize $\eta \le c_{\max}/L$*, and any* $T \ge \frac{1}{c_{\max}}\mathscr{T}$*, we have:*

$$\min\{f(\mathbf{u}_T) + g(\mathbf{u}_T) - f(\mathbf{u}_0) - g(\mathbf{u}_0), f(\mathbf{w}_T) + g(\mathbf{w}_T) - f(\mathbf{w}_0) - g(\mathbf{w}_0)\} \le -2.7\mathscr{F}$$

Proof. Without losing generality, let $\tilde{\mathbf{x}} = 0$ be the origin. Let $(c^{(2)}_{\max}, \hat{c})$ be the absolute constant so that Lemma 4 holds, also let $c^{(1)}_{\max}$ be the absolute constant to make Lemma 1 holds based on our current choice of $\hat{c}$. We choose $c_{\max} \le \min\{c^{(1)}_{\max}, c^{(2)}_{\max}\}$ so that our learning rate $\eta \le c_{\max}/L$ is small enough which make both Lemma 1 and Lemma 4 hold. Let $T^\star := \hat{c}\mathscr{T}$ and define:

$$T' = \inf_t \left\{t | \tilde{f}_{\mathbf{u}_0}(\mathbf{u}_t) + g(\mathbf{u}_t) - f(\mathbf{u}_0) - g(\mathbf{u}_0) \le -3\mathscr{F}\right\}$$

Let's consider following two cases:

Case $T' \le T^\star$*:* In this case, by Lemma 1, we know $\|\mathbf{u}_{T'-1}\| \le O(\mathscr{S})$, and therefore

$$\|\mathbf{u}_{T'}\| \le \|\mathbf{u}_{T'-1}\| + \eta\|\nabla f(\mathbf{u}_{T'-1})\| \le \|\mathbf{u}_{T'-1}\| + \eta\|\nabla f(\tilde{\mathbf{x}})\| + \eta L\|\mathbf{u}_{T'-1}\| \le O(\mathscr{S})$$

By choosing $c_{\max}$ small enough and $\eta \le c_{\max}/L$, this gives:

$$\begin{aligned}
&f(\mathbf{u}_{T'}) + g(\mathbf{u}_{T'}) - f(\mathbf{u}_0) - g(\mathbf{u}_0) \\
&\le \nabla f(\mathbf{u}_0)^\top(\mathbf{u}_{T'} - \mathbf{u}_0) + \frac{1}{2}(\mathbf{u}_{T'} - \mathbf{u}_0)^\top \nabla^2 f(\mathbf{u}_0)(\mathbf{u}_{T'} - \mathbf{u}_0) + \frac{\rho}{6}\|\mathbf{u}_{T'} - \mathbf{u}_0\|^3 + g(\mathbf{u}_{T'}) - g(\mathbf{u}_0) \\
&\le \tilde{f}_{\mathbf{u}_0}(\mathbf{u}_{T'}) - f(\mathbf{u}_0) + g(\mathbf{u}_{T'}) - g(\mathbf{u}_0) + \frac{\rho}{2}\|\mathbf{u}_0 - \tilde{\mathbf{x}}\|\|\mathbf{u}_{T'} - \mathbf{u}_0\|^2 + \frac{\rho}{6}\|\mathbf{u}_{T'} - \mathbf{u}_0\|^3 \\
&\le -3\mathscr{F} + O(\rho\mathscr{S}^3) = -3\mathscr{F} + O(\sqrt{\eta L} \cdot \mathscr{F}) \le -2.7\mathscr{F}
\end{aligned}$$

The first and second inequality exploit Hessian Lipschitz property of smooth function f, and $\|\mathbf{u}_0 - \tilde{\mathbf{x}}\| \le O(\mathscr{S})$, $\|\mathbf{u}_{T'} - \mathbf{u}_0\| \le O(\mathscr{S})$. By choose $c_{\max} \le \min\{1, \frac{1}{\hat{c}}\}$. We know $\eta < \frac{1}{L}$, by *sufficient decrease lemma* for proximal descent, we know each proximal descent iteration decreases function value. Therefore, for any $T \ge \frac{1}{c_{\max}}\mathscr{T} \ge \hat{c}\mathscr{T} = T^\star \ge T'$, we have:

$$\Phi(\mathbf{u}_T) - \Phi(\mathbf{u}_0) \le \Phi(\mathbf{u}_{T^\star}) - \Phi(\mathbf{u}_0) \le \Phi(\mathbf{u}_{T'}) - \Phi(\mathbf{u}_0) \le -2.7\mathscr{F}$$

Case $T' > T^\star$*:* In this case, by Lemma 1, we know $\|\mathbf{u}_t\| \le O(\mathscr{S})$ for all $t \le T^\star$. Define

$$T'' = \inf_t \left\{t | \tilde{f}_{\mathbf{w}_0}(\mathbf{w}_t) + g(\mathbf{w}_t) - f(\mathbf{w}_0) - g(\mathbf{w}_0) \le -3\mathscr{F}\right\}$$

By Lemma 4, we immediately have $T'' \le T^\star$. Apply same argument as in the case $T' \le T^\star$, we have for all $T \ge \frac{1}{c_{\max}}\mathscr{T}$ that $f(\mathbf{w}_T) + g(\mathbf{w}_T) - f(\mathbf{w}_0) - g(\mathbf{w}_0) \le f(\mathbf{w}_{T^\star}) + g(\mathbf{w}_{T^\star}) - f(\mathbf{w}_0) - g(\mathbf{w}_0) \le -2.7\mathscr{F}$.

3.5 Main Lemma

Lemma 6 (Main Lemma). *There exists universal constant $c_{\max}$, for $f(\cdot)$ satisfies A1, for any $\delta \in (0, \frac{d\kappa}{e}]$, suppose we start with point $\tilde{\mathbf{x}}$ satisfying following conditions:*

$$\|G(\tilde{\mathbf{x}})\| = \left\| L(\tilde{\mathbf{x}} - \text{prox}_{\frac{1}{L}g}\left(\tilde{\mathbf{x}} - \frac{1}{L}\nabla f(\tilde{\mathbf{x}})\right)\right\| \leq \mathscr{G} \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(\tilde{\mathbf{x}})) \leq -\gamma$$

Let $\mathbf{x}_0 = \tilde{\mathbf{x}} + \boldsymbol{\xi}$ where $\boldsymbol{\xi}$ come from the uniform distribution over ball with radius $\mathscr{S}/(\kappa \cdot \ln(\frac{d\kappa}{\delta}))$, and let $\mathbf{x}_t$ be the iterates of gradient descent from $\mathbf{x}_0$. Then, when stepsize $\eta \leq c_{\max}/L$, with at least probability $1-\delta$, we have following for any $T \geq \frac{1}{c_{\max}}\mathscr{T}$:

$$f(\mathbf{x}_T) + g(\mathbf{x}_T) - f(\tilde{\mathbf{x}}) - g(\tilde{\mathbf{x}}) \leq -\mathscr{F}$$

Proof. Denote $T_{\frac{l}{L}}(\mathbf{x}) = \text{prox}_{\frac{1}{L}g}\big[\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\big]$. The fisrt order stationary condition is equivalent to $\|\tilde{\mathbf{x}} - T_{\frac{1}{L}}(\tilde{\mathbf{x}})\| = \|\nabla f(\tilde{\mathbf{x}}) + \partial g\big(T_{\frac{1}{L}}(\tilde{\mathbf{x}})\big)\| \leq \mathscr{G}$, where ∂g is the subgradient of the function g.

As $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ has Lipschitz constant λ, we have

$$f(\mathbf{x}_0) + g(\mathbf{x}_0) \leq f(\tilde{\mathbf{x}}) + \langle \nabla f(\tilde{\mathbf{x}}), \boldsymbol{\xi}\rangle + \frac{L}{2}\|\boldsymbol{\xi}\|^2 + g(\tilde{\mathbf{x}}) + \langle \partial g(\tilde{\mathbf{x}}), \boldsymbol{\xi}\rangle + \frac{\lambda}{2}\|\boldsymbol{\xi}\|^2$$

Notice

$$\begin{aligned}\|\nabla f(\tilde{\mathbf{x}}) + \partial g(\tilde{\mathbf{x}})\| &= \|\nabla f(\tilde{\mathbf{x}}) + \partial g(T_{\frac{l}{L}}(\mathbf{x})) - \big(\partial g(T_{\frac{l}{L}}(\mathbf{x})) - \partial g(\tilde{\mathbf{x}})\big)\| \\ &\leq \mathscr{G} + \lambda\mathscr{G}\end{aligned}$$

By adding perturbation, in worst case we increase function value by:

$$\begin{aligned}f(\mathbf{x}_0) - f(\tilde{\mathbf{x}}) + g(\mathbf{x}_0) - g(\tilde{\mathbf{x}}) &\leq \|\nabla f(\tilde{\mathbf{x}}) + \partial g(\tilde{\mathbf{x}})\|\|\xi\| + \frac{L+\lambda}{2}\|\xi\|^2 \\ &\leq (1+\lambda)\mathscr{G}(\frac{\mathscr{S}}{\kappa \cdot \ln(\frac{d\kappa}{\delta})}) + \frac{1}{2}(L+\lambda)(\frac{\mathscr{S}}{\kappa \cdot \ln(\frac{d\kappa}{\delta})})^2 \\ &\leq (\frac{3}{2} + \frac{1}{5})\mathscr{F}\end{aligned}$$

where the last inequality follows from the fact that $\lambda \ll \min\{1, l\}$ per equation (7).

On the other hand, let radius $r = \frac{\mathscr{S}}{\kappa \cdot \ln(\frac{d\kappa}{\delta})}$. We know $\mathbf{x}_0$ come froms uniform distribution over $\mathbb{B}_{\tilde{\mathbf{x}}}(r)$. Let $\mathcal{X}_{\text{stuck}} \subset \mathbb{B}_{\tilde{\mathbf{x}}}(r)$ denote the set of bad starting points so that if $\mathbf{x}_0 \in \mathcal{X}_{\text{stuck}}$, then $\Phi(\mathbf{x}_T) - \Phi(\mathbf{x}_0) > -2.7\mathscr{F}$ (thus stuck at a saddle point); otherwise if $\mathbf{x}_0 \in B_{\tilde{\mathbf{x}}}(r) - \mathcal{X}_{\text{stuck}}$, we have $\Phi(\mathbf{x}_T) - \Phi(\mathbf{x}_0) \leq -2.7\mathscr{F}$.

By applying Lemma 5, we know for any $\mathbf{x}_0 \in \mathcal{X}_{\text{stuck}}$, it is guaranteed that $(\mathbf{x}_0 \pm \mu r\mathbf{e}_1) \notin \mathcal{X}_{\text{stuck}}$ where $\mu \in [\frac{\delta}{2\sqrt{d}}, 1]$. Denote $I_{\mathcal{X}_{\text{stuck}}}(\cdot)$ be the indicator function of being inside set $\mathcal{X}_{\text{stuck}}$; and vector $\mathbf{x} = (x^{(1)}, \mathbf{x}^{(-1)})$, where $x^{(1)}$ is the component along $\mathbf{e}_1$ direction, and $\mathbf{x}^{(-1)}$ is the remaining $d-1$ dimensional

vector. Recall $\mathbb{B}^{(d)}(r)$ be d-dimensional ball with radius r; By calculus, this gives an upper bound on the volumn of $\mathcal{X}_{\text{stuck}}$:

$$\begin{aligned}\text{Vol}(\mathcal{X}_{\text{stuck}}) &= \int_{\mathbb{B}^{(d)}_{\tilde{\mathbf{x}}}(r)} \mathrm{d}\mathbf{x} \cdot I_{\mathcal{X}_{\text{stuck}}}(\mathbf{x}) \\ &= \int_{\mathbb{B}^{(d-1)}_{\tilde{\mathbf{x}}}(r)} \mathrm{d}\mathbf{x}^{(-1)} \int_{\tilde{x}^{(1)}-\sqrt{r^2-\|\tilde{\mathbf{x}}^{(-1)}-\mathbf{x}^{(-1)}\|^2}}^{\tilde{x}^{(1)}+\sqrt{r^2-\|\tilde{\mathbf{x}}^{(-1)}-\mathbf{x}^{(-1)}\|^2}} \mathrm{d}x^{(1)} \cdot I_{\mathcal{X}_{\text{stuck}}}(\mathbf{x}) \\ &\le \int_{\mathbb{B}^{(d-1)}_{\tilde{\mathbf{x}}}(r)} \mathrm{d}\mathbf{x}^{(-1)} \cdot \left(2 \cdot \frac{\delta}{2\sqrt{d}} r\right) = \text{Vol}(\mathbb{B}_0^{(d-1)}(r)) \times \frac{\delta r}{\sqrt{d}}\end{aligned}$$

Then, we immediately have the ratio:

$$\frac{\text{Vol}(\mathcal{X}_{\text{stuck}})}{\text{Vol}(\mathbb{B}^{(d)}_{\tilde{\mathbf{x}}}(r))} \le \frac{\frac{\delta r}{\sqrt{d}} \times \text{Vol}(\mathbb{B}_0^{(d-1)}(r))}{\text{Vol}(\mathbb{B}_0^{(d)}(r))} = \frac{\delta}{\sqrt{\pi d}} \frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+\frac{1}{2})} \le \frac{\delta}{\sqrt{\pi d}} \cdot \sqrt{\frac{d}{2}+\frac{1}{2}} \le \delta$$

The second last inequality is by the property of Gamma function that $\frac{\Gamma(x+1)}{\Gamma(x+1/2)} < \sqrt{x+\frac{1}{2}}$ as long as $x \ge 0$. Therefore, with at least probability $1-\delta$, $\mathbf{x}_0 \notin \mathcal{X}_{\text{stuck}}$. In this case, we have:

$$\begin{aligned}\Phi(\mathbf{x}_T) - \Phi(\tilde{\mathbf{x}}) &= \Phi(\mathbf{x}_T) - \Phi(\mathbf{x}_0) + \Phi(\mathbf{x}_0) - \Phi(\tilde{\mathbf{x}}) \\ &\le -2.7\mathscr{F} + 1.7\mathscr{F} \le -\mathscr{F}\end{aligned}$$

which finishes the proof.

3.6 Main Theorem, and its Proof

Lemma 7 (Sufficient Decrease Lemma for Proximal Descent, [3]). *Assume the function f is real-valued and lower semi-continuous. Then for any $L \in (\frac{L}{2}, \infty)$ where $\eta = \frac{1}{L}$, we have $\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t+1}) \ge \frac{L-\frac{L}{2}}{L^2}\|G_{\frac{1}{L}}(\mathbf{x}_t)\|$.*

Proof of the Main Theorem

Proof. Denote $\tilde{c}_{\max}$ to be the absolute constant allowed in lemma 6 when it is given following parameters $\eta = \frac{c}{L}$, $\gamma = \sqrt{\rho\varepsilon}$, and $\delta = \frac{dL}{\sqrt{\rho\varepsilon}}\mathrm{e}^{-\chi}$. In this theorem, we let $c_{\max} = \min\{\tilde{c}_{\max}, 1/2\}$, and choose any constant $c \le c_{\max}$.

In this proof, we will actually achieve some point satisfying following condition:

$$\|G(\mathbf{x})\| \le g_{\text{thres}} \equiv \frac{\sqrt{c}}{\chi^2} \cdot \varepsilon, \qquad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge -\sqrt{\rho\varepsilon} \tag{19}$$

Since $c \le 1$, $\chi \ge 1$, we have $\frac{\sqrt{c}}{\chi^2} \le 1$, which implies any $\mathbf{x}$ satisfy Eq.(19) is also a ε-second-order stationary point.

Starting from $\mathbf{x}_0$, we know if $\mathbf{x}_0$ does not satisfy Eq.(19), there are only two possibilities:

1. $\|G(\mathbf{x}_0)\| > g_{\text{thres}}$: In this case, Algorithm 1 will not add perturbation. By lemma 7:
$$\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_0) \leq -\frac{\eta}{2} \cdot g_{\text{thres}}^2 = -\frac{c^2}{2\chi^4} \cdot \frac{\varepsilon^2}{L}$$
2. $\|G(\mathbf{x}_0)\| \leq g_{\text{thres}}$: In this case, Algorithm 1 will add a perturbation of radius r, and will perform proximal gradient descent (without perturbations) for the next t_{thres} steps. Algorithm 1 will then check termination condition. If the condition is not met, we must have:
$$\Phi(\mathbf{x}_{t_{\text{thres}}}) - \Phi(\mathbf{x}_0) \leq -\Phi_{\text{thres}} = -\frac{c}{\chi^3} \cdot \sqrt{\frac{\varepsilon^3}{\rho}}$$
This means on average every step decreases the function value by
$$\frac{\Phi(\mathbf{x}_{t_{\text{thres}}}) - \Phi(\mathbf{x}_0)}{t_{\text{thres}}} \leq -\frac{c^3}{\chi^4} \cdot \frac{\varepsilon^2}{L}$$

In case 1, we can repeat this argument for $t = 1$ and in case 2, we can repeat this argument for $t = t_{\text{thres}}$. Hence, we can conclude as long as algorithm 1 has not terminated yet, on average, every step decrease function value by at least $\frac{c^3}{\chi^4} \cdot \frac{\varepsilon^2}{L}$. However, we clearly can not decrease function value by more than $\Phi(\mathbf{x}_0) - \Phi^\star$, where $\Phi^\star$ is the function value of global minima. This means algorithm 1 must terminate within the following number of iterations:
$$\frac{\Phi(\mathbf{x}_0) - \Phi^\star}{\frac{c^3}{\chi^4} \cdot \frac{\varepsilon^2}{L}} = \frac{\chi^4}{c^3} \cdot \frac{L(\Phi(\mathbf{x}_0) - \Phi^\star)}{\varepsilon^2} = O\left(\frac{L(\Phi(\mathbf{x}_0) - \Phi^\star)}{\varepsilon^2} \ln^4\left(\frac{dL\Delta_\Phi}{\varepsilon^2\delta}\right)\right)$$

Finally, we would like to ensure when Algorithm 1 terminates, the point it finds is actually an ε-second-order stationary point. The algorithm can only terminate when the gradient mapping is small, and the function value does not decrease after a perturbation and t_{thres} iterations. We shall show every time when we add perturbation to iterate $\tilde{\mathbf{x}}_t$, if $\lambda_{\min}(\nabla^2 f(\tilde{\mathbf{x}}_t)) < -\sqrt{\rho\varepsilon}$, then we will have $\Phi(\mathbf{x}_{t+t_{\text{thres}}}) - \Phi(\tilde{\mathbf{x}}_t) \leq -\Phi_{\text{thres}}$. Thus, whenever the current point is not an ε-second-order stationary point, the algorithm cannot terminate.

According to Algorithm 1, we immediately know $\|G(\tilde{\mathbf{x}}_t)\| \leq g_{\text{thres}}$ (otherwise we will not add perturbation at time t). By lemma 6, we know this event happens with probability at least $1 - \frac{dL}{\sqrt{\rho\varepsilon}} e^{-\chi}$ each time. On the other hand, during one entire run of Algorithm 1, the number of times we add perturbations is at most:
$$\frac{1}{t_{\text{thres}}} \cdot \frac{\chi^4}{c^3} \cdot \frac{L(\Phi(\mathbf{x}_0) - \Phi^\star)}{\varepsilon^2} = \frac{\chi^3}{c} \frac{\sqrt{\rho\varepsilon}(\Phi(\mathbf{x}_0) - \Phi^\star)}{\varepsilon^2}$$

By the union bound, for all these perturbations, with high probability lemma 6 is satisfied. As a result Algorithm 1 works correctly. The probability of that is at least
$$1 - \frac{dL}{\sqrt{\rho\varepsilon}} e^{-\chi} \cdot \frac{\chi^3}{c} \frac{\sqrt{\rho\varepsilon}(\Phi(\mathbf{x}_0) - \Phi^\star)}{\varepsilon^2} = 1 - \frac{\chi^3 e^{-\chi}}{c} \cdot \frac{dL(\Phi(\mathbf{x}_0) - \Phi^\star)}{\varepsilon^2}$$

Recall our choice of $\chi = 3\max\{\ln(\frac{dL\Delta_f}{c\varepsilon^2\delta}), 4\}$. Since $\chi \geq 12$, we have $\chi^3 e^{-\chi} \leq e^{-\chi/3}$, this gives:

$$\frac{\chi^3 e^{-\chi}}{c} \cdot \frac{dL(\Phi(\mathbf{x}_0) - \Phi^\star)}{\varepsilon^2} \leq e^{-\chi/3} \frac{dL(\Phi(\mathbf{x}_0) - \Phi^\star)}{c\varepsilon^2} \leq \delta$$

which finishes the proof.

Remarks on large λ We point out that when λ is large enough so that the g term alters the local landscape of the objective function $\Phi(\mathbf{x})$, it is inevitable that new local minima will be introduced to the landscape of the objective function, and potentially change the stability of saddle points. We hypothesize that perturbed proximal descent will still converge to an ε-second-order stationary point regardless of the magnitude of λ.

An example for the new local minima introduced by large λ is Fig. 3b. We see new wrinkles are introduced to the four legs of the octopus function as λ increases from 1 to 10. If an iteration starts in the neighborhood of creases, it can converge to the bottom of the creases. Fig. 3c is an extreme scenario where the original landscape of the octopus function is completely altered to conform to the behavior of ℓ_1 penalty term.

3.7 From ε-second-order stationary point to local minimizers

Assumption A3 (Nondegenerate Saddle) *For all stationary points $\mathbf{x}_c$, $\exists\, m > 0$ such that $\min_{i=1,2,\cdots,d} |\lambda_i(\nabla^2 f(\mathbf{x}_c))| > m > 0$, where λ_i are the eigenvalues (not to be confused with the parameter λ).*

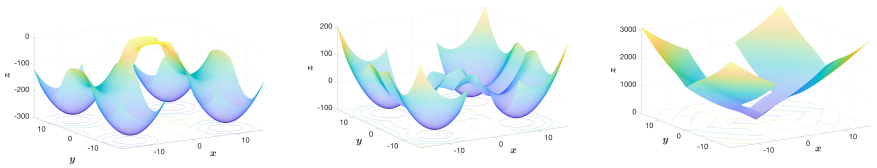
With this nondegenerate saddle assumption, the main theorem can be strengthened to the following corollary, whose proof is immediate as one sets the ε value in the main theorem as m^2/ρ and realizes that there is no eigenvalue of $\nabla^2 f$ existing between $-\sqrt{\rho\varepsilon}$ and the first positive eigenvalue.

Corollary 1. *There exists an absolute constant $c_{\max}$ such that if $f(\cdot)$ satisfies assumptions A1, A2 and A3, then for any $\delta > 0, \Delta_\Phi \geq \Phi(\mathbf{x}_0) - \Phi^\star$, constant $c \leq c_{\max}$, and $\varepsilon = \frac{m^2}{\rho}$, with probability $1-\delta$, the output of PPD$(\mathbf{x}_0, L, \rho, \varepsilon, c, \delta, \Delta_f)$ will be a local minimizer of $f + \lambda\|\mathbf{x}\|_1$, and terminate in iterations:*

$$\mathcal{O}\left(\frac{L(\Phi(\mathbf{x}_0) - \Phi^\star)}{\varepsilon^2} \ln^4\left(\frac{dL\Delta_\Phi}{\varepsilon^2\delta}\right)\right)$$

4 Numerical Experiment

We set f to be the "octopus" function described in [10] and use perturbed proximal descent to minimize the objective function $\Phi(\mathbf{x}) = f(\mathbf{x}) + \lambda\|\mathbf{x}\|_1$. Plots of octopus function defined in $\mathbb{R}^2$ for various λ are shown in Figure 3.



(a) $\lambda = 0.01$ (b) $\lambda = 10$ (c) $\lambda = 100$

Fig. 3: The octopus function with different λ values

The "octopus" family of functions is parameterized by τ, which controls the width of the "legs," and M and γ which characterize how sharp each side is surrounding a saddle point, related to the Lipschitz constant. The example illustrated in Fig. 3 uses parameters $M = \mathrm{e}, \gamma = 1, \tau = \mathrm{e}$.

We are interested in the octopus family of functions because it can be generalized to any dimension d, and it has $d-1$ saddle points (not counting the origin) which are known to slow down standard gradient descent algorithms. The usual minimization iteration sequence, if starting at the maximum value of the octopus function, will successively go through *each* saddle point before reaching the global minimum, thus rendering the iteration progress easy to track and visualize.

Specifics of Octopus Function Define the *auxiliary gluing functions* as

$$\mathcal{G}_1(x_i) = -\gamma x_i^2 + \frac{-14L + 10\gamma}{3\tau}(x_i - \tau)^3 + \frac{5L - 3\gamma}{2\tau}(x_i - \tau)^4$$
$$\mathcal{G}_2(x_i) = -\gamma - \frac{10(L+\gamma)}{\tau^3}(x_i - 2\tau)^3 - \frac{15(L+\gamma)}{\tau^4}(x_i - 2\tau)^4 - \frac{6(L+\gamma)}{\tau^5}(x_i - 2\tau)^5$$

Define the *gluing function* and *gluing balance constant* respectively as

$$\mathcal{G}(x_i, x_{i+1}) = \mathcal{G}_1(x_i) + \mathcal{G}_2(x_i)x_{i+1}^2$$
$$\nu = -\mathcal{G}_1(2\tau) + 4L\tau^2 = \frac{26L + 2\gamma}{3}\tau^2 + \frac{-5L + 3\gamma}{2}\tau^3$$

For a given $i = 1, \cdots, d-1$, when $6\tau \geq x_1, \cdots, x_{i-1} \geq 2\tau, \tau \geq x_i \geq 0, \tau \geq x_{i+1}, \cdots, x_d \geq 0$

$$f(\mathbf{x}) = \sum_{j=1}^{i-1} L(x_j - 4\tau)^2 - \gamma x_i^2 + \sum_{j=i+1}^{d} Lx_j^2 - (i-1)\nu \equiv f_{i,1}(\mathbf{x}) \tag{20}$$

and if $6\tau \geq x_1, \cdots, x_{i-1} \geq 2\tau, 2\tau \geq x_i \geq \tau, \tau \geq x_{i+1}, \cdots, x_d \geq 0$, we have

$$f(\mathbf{x}) = \sum_{j=1}^{i-1} L(x_j - 4\tau)^2 + \mathcal{G}(x_i, x_{i+1}) + \sum_{j=i+2}^{d} Lx_j^2 - (i-1)\nu \equiv f_{i,2}(\mathbf{x}) \tag{21}$$

and for $i = d$, if $6\tau \geq x_1, \cdots, x_{d-1} \geq 2\tau, \tau \geq x_d \geq 0$

$$f(\mathbf{x}) = \sum_{j=1}^{d-1} L(x_j - 4\tau)^2 - \gamma x_d^2 - (d-1)\nu \equiv f_{d,1}(\mathbf{x}) \tag{22}$$

and if $6\tau \geq x_1, \cdots, x_{d-1} \geq 2\tau, 2\tau \geq x_d \geq \tau$

$$f(\mathbf{x}) = \sum_{j=1}^{d-1} L(x_j - 4\tau)^2 + \mathcal{G}_1(x_d) - (d-1)\nu \equiv f_{d,2}(\mathbf{x}) \tag{23}$$

and if $6\tau \geq x_1, \cdots, x_d \geq 2\tau$,

$$f(\mathbf{x}) = \sum_{j=1}^{d} L(x_j - 4\tau)^2 - d\nu \equiv f_{d+1,1}(\mathbf{x}) \tag{24}$$

Remark All saddle points happen at $(\pm 4\tau, \pm 4\tau, \cdots, \pm 4\tau, 0, 0, \cdots, 0)$, and the global minimum is at $(\pm 4\tau, \cdots, \pm 4\tau)$. Regions in the form of $[2\tau, 6\tau] \times \cdots \times [2\tau, 6\tau] \times [\tau, 2\tau] \times [0, \tau] \times \cdots \times [0, \tau]$ are transition zones described by the gluing functions which connect separate pieces to make f a continuous function. The octopus function can be constructed first in the first quadrant, and then using even function reflection to define it in the all other quadrants. A typical descent algorithm applied to the octopus generates iterations that take multiple turns like walking down a spiral staircase, each staircase leading to a new dimension.

4.1 Results

We apply the perturbed proximal descent (PPD) on the octopus function plus $0.01\|\mathbf{x}\|_1$ when the dimension varies between $d = 2, 5, 10, 20$. We set the constant $c = 3$. For comparison, we apply perturbed gradient descent (PGD) as well since $\|\mathbf{x}\|_1$ is differentiable almost everywhere; for both algorithms, the norm of the perturbation $\boldsymbol{\xi}$ is 0.1.

We see that PPD successfully finds the local minimum in the first three cases within 1000 iterations, and in the case of $d = 20$, PPD almost finds the local minimum within 1000 iterations. In contrast, unperturbed proximal descent (PD), gradient descent (GD), and perturbed gradient descent (PGD) sequences are trapped near saddle points.

5 Conclusion

This paper provides an algorithm to minimize a non-convex function plus a ℓ_1 penalty of small magnitude, with a probabilistic guarantee that the returned result is an approximate second-order stationary point, and hence for a large

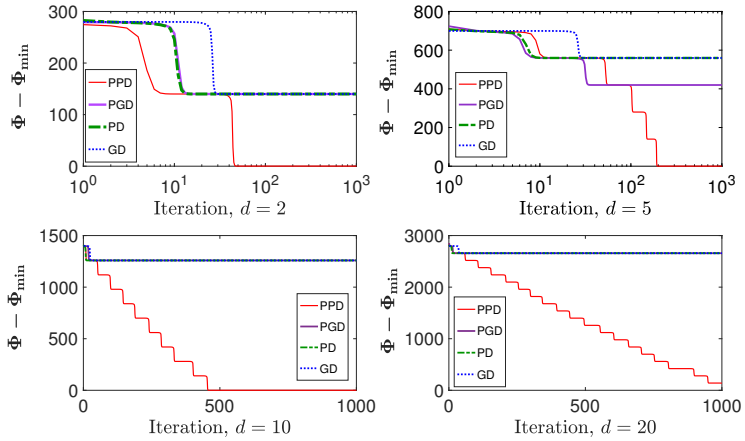


Fig. 4: Performance of our proposed PPD algorithm on the octopus function with $\lambda = 0.01$

class of functions, a local minimum instead of a saddle point. The complexity is of $\mathcal{O}(\varepsilon^{-2})$ and the result depends on dimension in $\mathcal{O}(\ln^4 d)$.

The deficiency of the result is that the magnitude of ℓ_1 penalty needs to be small to let our theoretical result hold. Meanwhile, we also notice that a large λ will lead to creation of new local minima to the objective function altering the original landscape. Our future work will address the case of large λ in the iteration process.

References

1. H. Attouch, J. Bolte, and B.F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss-Seidel methods. Mathematical Programming, pages 1–39, 2011.
2. H. H. Bauschke and P. L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer-Verlag, New York, 2 edition, 2017.
3. A. Beck. First-Order Methods in Optimization. MOS-SIAM Series on Optimization, 2017.
4. J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Prog., 146(1-2):459–494, 2014.
5. R.I. Bot, E.R.. Csetnek, and D-K Nguyen. A proximal minimization algorithm for structured nonconvex and nonsmooth problems. arXiv preprint arXiv:1805.11056v1[math.OC], 2018.
6. Y. Carmon, J. Duchi, O. Hinder, and A. Sidford. Accelerated methods for nonconvex optimization. SIAM Journal on Optimization, 28(2):1751–1772, 2018.
7. P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. SIAM Multiscale Model. Simul., 4(4):1168–1200, 2005.

8. F.E. Curtis, D.P. Robinson, and M. Samadi. A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{\frac{3}{2}})$ for nonconvex optimization. Mathematical Programming, 162(1):1–32, Mar 2017.
9. Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Advances in Neural Information Processing Systems, pages 2933–2941, 2014.
10. Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In Advances in Neural Information Processing Systems, pages 1067–1077, 2017.
11. F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. Neural computation, 7(2):219–269, 1995.
12. S. J. Reddi, S. Sra, B. Poczos, and A.J. Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 1145–1153. Curran Associates, Inc., 2016.
13. C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In ICML, 2017.
14. J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In Conference on Learning Theory, pages 1246–1257, 2016.
15. Y. Liu and W. Yin. An envelope for Davis-Yin splitting and strict saddle point avoidance. arXiv preprint arXiv:1804.08739, 2018.
16. Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. Doklady AN SSSR, translated as Soviet Math. Docl., 269:543–547, 1983.
17. Yurii Nesterov and Boris T. Polyak. Cubic regularization of newton method and its global performance. Math. Program., 108:177–205, 2006.
18. N. Z. Shor. An application of the method of gradient descent to the solution of the network transportation problem. Materialy Naucnovo Seminara po Teoret i Priklad. Voprosam Kibernet. i Issted. Operacii, Nucnyi Sov. po Kibernet, Akad. Nauk Ukrain. SSSR, vyp, 1:9–17, 1962.
19. L. Stella, A. Themelis, and P. Patrinos. Forward-backward quasi-Newton methods for nonsmooth optimization problems. Computational Optimization and Applications, 67(3):443–487, 2017.
20. Y. Xu, R. Jin, and T. Yang. First-order stochastic algorithnms for escaping from saddle points in almost linear time. arXiv preprint, 2018. arXiv:1711.01944v3 [math.OC].
21. Z. Zhu and Y. Li. Neon2: Finding local minima via first-order oracles. arXiv preprint, 2018. arXiv:1711.06673 [cs.LG].