



## Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:  
<http://pubsonline.informs.org>

### One-Step Estimation with Scaled Proximal Methods

Robert Bassett, Julio Deride

To cite this article:

Robert Bassett, Julio Deride (2022) One-Step Estimation with Scaled Proximal Methods. Mathematics of Operations Research 47(3):2366-2386. <https://doi.org/10.1287/moor.2021.1212>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact [permissions@informs.org](mailto:permissions@informs.org).

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2021, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

# One-Step Estimation with Scaled Proximal Methods

Robert Bassett,<sup>a</sup> Julio Deride<sup>b</sup><sup>a</sup>Department of Operations Research, Naval Postgraduate School, Monterey, California 93943; <sup>b</sup>Department of Mathematics, Universidad Técnica Federico Santa María, Valparaíso 8940000, ChileContact: robert.bassett@nps.edu,  <https://orcid.org/0000-0002-8056-4417> (RB); julio.deride@usm.cl, <https://orcid.org/0000-0003-4094-8819> (JD)

Received: January 19, 2021

Revised: June 23, 2021; July 30, 2021

Accepted: September 6, 2021

Published Online in Articles in Advance:  
December 13, 2021MSC2020 Subject Classification: Primary:  
62F12; secondary: 65K10, 90C30<https://doi.org/10.1287/moor.2021.1212>

Copyright: © 2021 INFORMS

**Abstract.** We study statistical estimators computed using iterative optimization methods that are not run until completion. Classical results on maximum likelihood estimators (MLEs) assert that a *one-step estimator* (OSE), in which a single Newton-Raphson iteration is performed from a starting point with certain properties, is asymptotically equivalent to the MLE. We further develop these early-stopping results by deriving properties of one-step estimators defined by a single iteration of scaled proximal methods. Our main results show the asymptotic equivalence of the likelihood-based estimator and various one-step estimators defined by scaled proximal methods. By interpreting OSEs as the last of a sequence of iterates, our results provide insight on scaling numerical tolerance with sample size. Our setting contains scaled proximal gradient descent applied to certain composite models as a special case, making our results applicable to many problems of practical interest. Additionally, our results provide support for the utility of the scaled Moreau envelope as a statistical smoother by interpreting scaled proximal descent as a quasi-Newton method applied to the scaled Moreau envelope.

**Funding:** The work of R. Bassett was supported by the Office of Naval Research Global [Grants N0001419WX00183 and N0001420WX01523]. The work of J. Deride was funded by the National Agency for Research and Development [Grant FONDECYT 11190549].

**Keywords:** proximal operator • one-step estimator • Moreau envelope • proximal-gradient

## 1. Introduction

In likelihood-based statistical inference, estimators are defined as solutions to optimization problems, with the objective function constructed from a random sample. When computing the estimator, however, it is often the case that the statistical context for the optimization problem is disregarded and a purely numerical perspective adopted. Numerical tolerance is taken to be as small as can be computed in a reasonable amount of time, and the resulting approximate minimizer is taken as the realization of the estimator for the given data sample.

In this paper, we view statistical and numerical error holistically. We retain the statistical origin of the mathematical program defining the estimator in order to provide insight on the numerical tolerance required to achieve statistical optimality. By *statistical optimality*, we mean asymptotic equivalence to the defined estimator, a minimizer which depends on a random sample. For reasons that will become clear in Section 5, we focus our attention on scaled methods. We consider scaled proximal gradient descent and scaled proximal descent, and generalizations of proximal gradient descent and proximal descent, respectively, which adjust for curvature of the objective function. We show that one-step estimators constructed from these algorithms achieve statistical optimality in an asymptotic sense for a broad class of parametric families and likelihood-based estimators.

The rest of this paper is organized as follows. In Section 1, we review one-step estimation for statistical inference in parametric models and summarize previous work on scaled proximal algorithms. In Section 2, we give one-step estimation results for scaled proximal gradient descent applied to a composite model, such as is commonly encountered in penalized and constrained M-estimation. Section 3 contains similar results for scaled proximal descent. The scaled and unscaled proximal operator can be interpreted as scaled and unscaled gradient descent, respectively, applied to an inf-convolution of the objective function, so in Section 4 we provide an alternative interpretation of our results using variational analysis to describe the scaled Moreau envelope as statistical smoother. We conclude with Section 5, which provides a counterexample demonstrating that one-step estimation results of the type we consider do not hold for first-order methods. In this section, we also provide numerical validation of our results and some examples demonstrating their utility.

### 1.1. One-Step Estimation

It is well known that, for a general class of statistical models, the MLE is asymptotically unbiased and efficient, meaning that as the number of samples goes to infinity, the expectation of the MLE matches the parameter to be

estimated and its variance attains the Cramer-Rao lower bound. We begin by reviewing aspects of this theory relevant to our contributions.

Let  $\Theta \subseteq \mathbb{R}^d$  be an open set, and let  $\{P_\theta : \theta \in \Theta\}$  be a parametric family of probability measures on a measurable space  $(\mathcal{X}, \mathcal{A}, \mu)$ , where  $\mu$  is  $\sigma$ -finite. Assume that for all  $\theta$  the measure  $P_\theta$  is absolutely continuous with respect to  $\mu$  and hence has Radon-Nikodym derivative  $p_\theta := dP_\theta/d\mu$ , and its support does not depend on  $\theta$ . Fix  $\theta^* \in \Theta$ , and assume that  $X_1, \dots, X_n \sim \text{i.i.d. } p_{\theta^*}$ . We will exclusively focus on estimators  $\hat{\theta}$  that converge at the  $\sqrt{n}$  parametric rate, so that  $\sqrt{n}(\hat{\theta} - \theta^*)$  converges in law to some limit distribution.

A sequence of random variables  $X_n$  is stochastically bounded, denoted by  $O_P(1)$ , if  $X_n$  is bounded in probability with respect to the measure  $P_{\theta^*}$ . Additionally,  $X_n$  is  $o_P(1)$  if it converges to zero in probability with respect to  $P_{\theta^*}$ . An estimator  $\hat{\theta}$  of  $\theta_0$  is said to be  $\tau_n$ -consistent, for a given sequence  $\tau_n$ , if  $\tau_n(\hat{\theta} - \theta_0)$  is  $O_P(1)$ . We also call  $\hat{\theta}$   $\tau_n$ -consistent for another estimator  $\tilde{\theta}$  if  $\tau_n(\hat{\theta} - \tilde{\theta}) = O_P(1)$ . Because we only consider estimators that converge at the  $\sqrt{n}$  parametric rate, two estimators  $\hat{\theta}$  and  $\tilde{\theta}_2$  are asymptotically equivalent when they are  $\sqrt{n}$ -consistent.

The statistical model  $\{P_\theta : \theta \in \Theta\}$  is said to be differentiable in quadratic mean at some fixed  $\theta_0 \in \Theta$  if there exists a measurable vector-valued function  $\nabla \ell_{\theta_0}$  such that, as  $\theta \rightarrow \theta_0$ ,

$$\int \left[ \sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^\top \nabla \ell_{\theta_0} \sqrt{p_{\theta_0}} \right]^2 d\mu = o(\|\theta - \theta_0\|^2).$$

Differentiability in quadratic mean is a relaxed smoothness assumption that still permits many of the classical results related to maximum likelihood. For example, it can be shown that a location model with univariate Laplace density, where  $p_\theta = \frac{1}{2}e^{-|x-\theta|}$  for  $\theta \in \mathbb{R}$ , is differentiable in quadratic mean with  $\nabla \ell_\theta(x) = \text{sign}(x - \theta)$ , even though it is nonsmooth. For a statistical model that is differentiable in quadratic mean, the Fisher Information at  $\theta_0$  is defined at  $I_{\theta_0} = E[\nabla \ell_{\theta_0} \nabla \ell_{\theta_0}^\top]$ . The following theorem, due to Le Cam [25], gives the asymptotic properties of the maximum likelihood estimator; details can be found in Van der Vaart [42].

**Theorem 1.** *Suppose that the model  $\{P_\theta : \theta \in \Theta\}$  is differentiable in quadratic mean at  $\theta^* \in \text{int}(\Theta)$ . Suppose also that there exists a measurable function  $L$  with  $E[L^2(X_1)]$  finite and such that, for every neighborhood of  $\theta^*$  and every  $\theta_1$  and  $\theta_2$  in that neighborhood,*

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq L(x)\|\theta_1 - \theta_2\|, \text{ for every } x \text{ in } \mathbb{R}^d.$$

*If the Fisher information matrix  $I_{\theta^*}$  is nonsingular and the MLE  $\hat{\theta}_{mle}$  is consistent, then  $\sqrt{n}(\hat{\theta}_{mle} - \theta^*)$  is asymptotically normal with mean zero and covariance matrix  $I_{\theta^*}^{-1}$ .*

Theorem 1 gives that the MLE is statistically optimal in a specific sense. According to the Cramer-Rao bound, an unbiased estimator  $\hat{\theta}$  of  $\theta^*$  must satisfy the matrix inequality  $\text{Var}(\hat{\theta}) \geq I_{\theta^*}^{-1}$ . Because  $\hat{\theta}_{mle}$  attains this variance bound as  $n \rightarrow \infty$ , the MLE is optimal in terms of its asymptotic bias and variance. Moreover, this optimality of the MLE is the primary property that motivates its use. According to Van der Vaart [42, p. 64], “The justification through asymptotics appears to be the only general justification of the method of maximum likelihood.” (See Le Cam [27] for a similar opinion.) Though there are many important contributions related to finite-sample results for maximum likelihood estimation (Boucheron and Massart [9], Spokoiny [39]), they apply only with sub-Gaussian or subexponential tail behavior. Our contributions are intended to apply to likelihood-based inference generally, so we will use asymptotic justifications for our results.

Asymptotic efficiency can be considered in the more general setting of M-estimation as well. For each  $\theta \in \Theta$ , let  $m_\theta : \mathcal{X} \rightarrow \mathbb{R}$  be a measurable function. The M-estimator for this criterion is defined as

$$\hat{\theta}_M = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n m_\theta(X_i).$$

If  $m_\theta = -\log p_\theta$ , then the corresponding M-estimator coincides with the maximum likelihood estimator. M-estimation has an asymptotic result similar to Theorem 1 when  $E[m_\theta(X)]$  has a second-order expansion (Van der Vaart [42, theorem 5.23]). Interestingly, when applying this theorem in the context of maximum likelihood, differentiability in quadratic mean provides the existence of this second-order Taylor expansion. It is remarkable that a first-order assumption can be used to provide second-order properties; we refer the reader to Pollard [35] for further discussion on this point.

MLEs and M-estimators are often computed with iterative methods, so it is important to quantify the performance of estimators derived as iterates of an optimization method. One approach for doing so is one-step

estimation, which was first introduced by Le Cam [24]. In its original form, one-step estimation consisted of applying one Newton-Raphson iteration, on a maximum log-likelihood objective, to an initial estimator within some range of the parameter to be estimated. Le Cam showed that, subject to certain conditions, this one-step estimator is asymptotically equivalent to the MLE. This is remarkable because it shows that the result of a single Newton-Raphson iteration possesses the same statistical optimality as the limit point of Newton-Raphson iterates, the MLE. One-step estimation has since been extended to M-estimation (Bickel [8]), sparse estimation (Taddy [40], Zou [44]), quasilielihood estimation (Fan and Chen [14]), and distributed computation (Huang and Huo [22]).

The one-step estimation result most relevant to our work is the following theorem, which summarizes the asymptotic performance of a certain one-step estimator.

**Theorem 2** (Van der Vaart [42, theorems 5.21 and 5.45]). *Let  $\hat{\theta}^*$  be an M-estimator with a continuously differentiable criterion  $m_\theta$ , where  $E[\|\nabla m_{\theta^*}\|^2] < \infty$ ,  $E[\nabla m_{\theta^*}] = 0$ ,  $E[m_\theta]$  is twice differentiable at  $\theta^*$  with nonsingular Hessian  $V_{\theta^*}$ , and  $\frac{1}{n} \sum_{i=1}^n \nabla m_{\hat{\theta}_M}(X_i) = o_P(n^{-1/2})$ . Assume that  $\hat{\theta}_M$  is consistent for  $\theta^*$  and that there is a measurable function  $L$  with  $E[L^2] < \infty$  such that, for every  $\theta_1$  and  $\theta_2$  in a neighborhood of  $\theta^*$ ,*

$$\|\nabla m_{\theta_1}(x) - \nabla m_{\theta_2}(x)\| \leq L(x) \|\theta_1 - \theta_2\|.$$

*Let  $\hat{\theta}_{init}$  be a  $\sqrt{n}$ -consistent estimator of  $\theta^*$ . If  $C_n$  is a sequence of random matrices such that  $C_n \rightarrow^P V_{\theta^*}$ , then  $\hat{\theta}_M$  and  $\hat{\theta}_{init} - C_n^{-1} \left( \frac{1}{n} \sum_{i=1}^n \nabla m_{\hat{\theta}_{init}}(X_i) \right)$  are asymptotically equivalent.*

This result on one-step estimation gives that, subject to some mild conditions, one Newton step performed on an initial estimator within  $n^{-1/2}$  of  $\hat{\theta}_M$  gives the same large sample performance as the M-estimator itself.

Though one-step estimation is defined as a single step of an iterative method, it can be interpreted as the last step of an iterative method. As long as the penultimate iterate satisfies the properties required of  $\hat{\theta}_{init}$ , the last iteration generates an estimator with one-step properties. Moreover, further iterations do not increase the performance of the estimator from the asymptotic perspective. Theorem 2 demonstrates that the one-step estimator is asymptotically equivalent to  $\hat{\theta}_M$ ; hence it is  $\sqrt{n}$ -consistent. The one-step estimator then can be used as  $\hat{\theta}_{init}$  in another iteration, because it satisfies the required properties for the starting point, but the resulting asymptotic distribution remains the same. We conclude that numerical tolerance for Newton’s method should be  $O(n^{-1/2})$ , where  $n$  is the sample size of the problem, in order to respect the statistical origin of the problem and guarantee the asymptotic properties in Theorem 2.

### 1.2. Proximal Methods

In this section, we review proximal descent and proximal gradient descent, in addition to their extensions scaled proximal descent and proximal Newton descent. Each of these algorithms are iterative methods for finding local minima, and in each step these methods use a proximal operator, a numerical primitive that involves solving a related optimization problem.

For a function  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  and positive parameter  $\lambda > 0$ , the proximal operator  $\text{prox}_f^\lambda(x)$  is defined as

$$\text{prox}_f^\lambda(x) = \arg \min_{w \in \mathbb{R}^d} \left\{ f(w) + \frac{1}{2\lambda} \|w - x\|^2 \right\}.$$

The proximal operator is a generalization of a projection. Indeed, if  $S \subset \mathbb{R}^d$  and  $I_S$  is an indicator function on the set  $S$ ,

$$I_S(x) = \begin{cases} 0 & \text{if } x \in S \\ \infty & \text{otherwise,} \end{cases}$$

then the proximal operator of  $I_S$  for any  $\lambda > 0$  is  $\arg \min_{w \in S} \|w - x\|^2$ , the projection onto the set  $S$ . Alternatively, one can view the proximal operator as a kind of implicit gradient descent. If  $f$  is smooth and convex, then the optimality conditions give

$$x - \lambda \nabla f \left( \text{prox}_f^\lambda(x) \right) = \text{prox}_f^\lambda(x).$$

Thus, applying the proximal operator is equivalent to taking a gradient step, where the gradient is evaluated at the output of the proximal operator instead of the initial point  $x$ , and the step length is  $\lambda$ .

If the objective function  $f$  is proper, so that it is finite for at least one point and never takes the value  $-\infty$ , then  $x^*$  minimizing  $f$  implies  $\text{prox}_f^\lambda(x^*) = x^*$ . Because the proximal operator is also firmly nonexpansive (Bauschke and Combettes [3, proposition 12.28]) when  $f$  is convex, so that  $\|\text{prox}_f^\lambda(x) - \text{prox}_f^\lambda(y)\| \leq \|x - y\|$ , a natural approach to minimizing a convex  $f$  is to iterate the proximal operator (Bauschke and Combettes [3, proposition 12.29]). This

algorithm is called *proximal descent*, and it converges for any sequence  $\lambda_k$  of the  $\lambda$  parameters such that  $\lambda > 0$  and  $\sum_k \lambda_k \rightarrow \infty$ . Proximal descent is notable for its simplicity, but it finds limited application in practice, because its convergence rate matches gradient descent (Beck [4]). Evaluation of the proximal operator is rarely as simple as gradient evaluation, so besides a few notable exceptions (Golub and Wilkinson [18], Mattingley and Boyd [31]), proximal descent is primarily of theoretical interest. Additional details of the proximal operator and the proximal descent method can be found in (Bauschke and Combettes [3, sections 12.4 and 28.5]).

Other methods based on proximal descent are extremely useful in practice. Consider the composite model

$$\min_{x \in \mathbb{R}^d} g(x) + h(x),$$

where  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  is closed and continuously differentiable and  $h : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  is convex and potentially non-smooth. If a point  $x^*$  minimizes  $g(x) + h(x)$ , then it can easily be shown that<sup>1</sup>

$$0 \in \nabla g(x^*) + \partial h(x^*),$$

where  $\partial$  denotes the subdifferential of a convex function. A few algebraic manipulations give that

$$x^* - \lambda \nabla g(x^*) \in x^* + \lambda \partial h(x^*).$$

Hence we conclude that  $x^* \in \text{prox}_h^\lambda(x^* - \lambda \nabla g(x^*))$ . If, in addition,  $g$  is a strongly convex function with  $L$ -Lipschitz continuous gradient, then the operator  $\text{prox}_h^\lambda(x - \lambda \nabla g(x))$  is firmly nonexpansive for values of  $\lambda$  not larger than  $\frac{1}{L}$  (Parikh and Boyd [34]). Thus, we iterate  $\text{prox}_h^\lambda(x - \lambda \nabla g(x))$  in order to find a minimizer of  $g(x) + h(x)$ . Because each step of this algorithm composes a proximal step in  $h$  with a gradient descent step in  $g$ , this method is called *proximal gradient descent*. Proximal gradient descent finds widespread application in many problems in statistics and machine learning, where the strongly convex function  $g$  is most often a data fidelity term and  $h$  is a penalty or constraint that encourages a certain solution structure (Polson et al. [36]). In the setting of Bayesian point estimation, the function  $h$  can be used to incorporate a prior using maximum a posteriori (MAP) estimation (Bassett and Deride [2]).

As alternatives to proximal descent and proximal gradient descent, we consider their scaled analogues. For a function  $f$  and positive definite matrix  $C$ , we define the *scaled proximal operator* as

$$\text{prox}_f^C(x) := \arg \min_{w \in \mathbb{R}^d} \left\{ f(w) + \frac{1}{2} \|w - x\|_C^2 \right\},$$

where  $\|\cdot\|_C$  is the scaled norm induced by  $C$ ; that is,  $\|x\|_C^2 = \langle x, Cx \rangle$  for every  $x$  in  $\mathbb{R}^d$ . By replacing the proximal operator with the scaled proximal operator in proximal descent, we have scaled proximal descent. Similarly, replacing the step length  $\lambda$  in proximal gradient descent with a scaling matrix  $C^{-1}$  yields scaled proximal gradient descent. The only notable difference is that the firm nonexpansivity in both cases is now with respect to the  $\|\cdot\|_C$  norm. Various conventions to generate  $C$  have been previously investigated, including scalar multiples of the identity (Beck and Teboulle [5], Milzarek [32]), diagonal matrices (Tseng and Yun [41]), quasi-Newton approximations of  $\nabla^2 f$  (Becker and Fadili [7], Kanzow and Lechner [23], Scheinberg and Tang [38]), and taking  $C = \nabla^2 f$  (Lee et al. [28]). Better approximations of the curvature of  $f$  generally make the scaled proximal operator harder to evaluate but also result in scaled proximal descent requiring fewer iterations to reach a fixed solution accuracy (Friedlander and Goh [16]).

## 2. The Composite Model

A *composite model* refers to an optimization problem  $\min f$ , where  $f = g + h$ . It is common practice to assume that  $g$  is convex with Lipschitz continuous gradient, whereas the function  $h$  is convex and potentially nonsmooth. Where possible, we will also relax the convexity assumption on  $g$ .

In this section, we study the asymptotic equivalence of an estimator defined as the minimizer in a composite model and the one-step estimator attained through scaled proximal gradient descent. Let  $g_n : \Omega \times \Theta \rightarrow \mathbb{R}$  and  $h_n : \Omega \times \Theta \rightarrow \mathbb{R} \cup \{\infty\}$  be sequences of random loss functions, where  $\Theta \subseteq \mathbb{R}^d$  and  $\Omega$  forms some probability space  $(\Omega, \mathcal{A}, P)$ . More precisely, we mean that  $g_n(\cdot, \theta)$  and  $h_n(\cdot, \theta)$  are measurable functions on  $\Omega$  for each  $\theta \in \Theta$ . Similarly,  $g_n(\omega, \cdot)$  and  $h_n(\omega, \cdot)$  define loss functions on  $\Theta$  for each  $\omega \in \Omega$ . Let  $C_n$  be a sequence of  $d \times d$  random matrices defined on the same probability space. Note that this setting is more general than the parametric estimation from independent and identically distributed (i.i.d.) observations introduced in Section 1.1. We will return to the i.i.d. setting in Proposition 1.

We suppress the dependence on  $\Omega$  in our notation, so that  $g_n(\theta)$  and  $h_n(\theta)$  denote a sequence of random variables indexed by  $\theta \in \Theta$ . Similarly, gradients of  $g_n$  and subgradients of  $h_n$  are with respect to the  $\Theta$  argument. Define an estimator for the composite model  $g_n + h_n$  as

$$\hat{\theta} = \arg \min_{\theta \in \Theta} g_n + h_n, \tag{1}$$

which is assumed to exist  $P$ -almost surely. Note that  $\hat{\theta}$  is a function of  $n$ , though to ease notation we have omitted this dependence.

For an initial estimator  $\hat{\theta}_{init}$  of  $\hat{\theta}$ , define the associated one-step estimator (OSE), denoted by  $\hat{\theta}_{ose}$ , as one iteration of scaled proximal gradient descent from the initial estimator  $\hat{\theta}_{init}$ ,

$$\hat{\theta}_{ose} = \text{prox}_{h_n}^{C_n}(\hat{\theta}_{init} - C_n^{-1} \nabla g_n(\hat{\theta}_{init})), \tag{2}$$

which is also assumed to exist  $P$ -almost surely. The initial estimator  $\hat{\theta}_{init}$  is assumed to be a  $\sqrt{n}$ -consistent estimator of  $\hat{\theta}$ , so that  $\sqrt{n}(\hat{\theta}_{init} - \hat{\theta})$  is bounded in probability.

Our first result provides conditions under which the estimators in Equations (1) and (2) are asymptotically equivalent, so that the one-step estimator has the same asymptotic performance as the minimizer of the composite model.

**Theorem 3.** Consider the composite estimation model in Equation (1), where it holds with probability 1 that  $g_n$  is continuously differentiable and  $h_n$  is convex and proper.

Assume the following.

(i) For each  $M > 0$ , there is a positive definite matrix  $H$  such that

$$\sup_{\sqrt{n} \|\theta - \hat{\theta}\| < M} \sqrt{n} (\nabla g_n(\theta) - \nabla g_n(\hat{\theta})) - \sqrt{n} H(\theta - \hat{\theta}) \rightarrow^P 0.$$

(ii) We have that  $C_n > 0$  and  $C_n^{-1} H \rightarrow^P I$ , where  $I$  denotes the  $d \times d$  identity matrix.

Then for any  $\sqrt{n}$ -consistent estimator,  $\hat{\theta}_{init}$  of  $\hat{\theta}$ ,  $\hat{\theta}$  and the one-step estimator  $\hat{\theta}_{ose}$  of Equation (2) are asymptotically equivalent.

**Proof.** We first show that for some constant  $K$ , the scaled proximal operator  $\text{prox}_{h_n}^{C_n}$  is  $K$ -Lipschitz continuous with high probability. Indeed, for any  $\theta$  the proximal operator satisfies

$$C_n \left( \theta - \text{prox}_{h_n}^{C_n}(\theta) \right) \in \partial h \left( \text{prox}_{h_n}^{C_n}(\theta) \right).$$

So, by the monotonicity of the subdifferential of  $h_n$ , for any  $\theta_1, \theta_2$  we have

$$\left( \text{prox}_{h_n}^{C_n}(\theta_1) - \text{prox}_{h_n}^{C_n}(\theta_2) \right)^T C_n \left( \theta_1 - \text{prox}_{h_n}^{C_n}(\theta_1) - \left( \theta_2 - \text{prox}_{h_n}^{C_n}(\theta_2) \right) \right) \geq 0.$$

Applying the Cauchy-Schwarz inequality gives

$$\|\theta_1 - \theta_2\|_{C_n} \geq \|\text{prox}_{h_n}^{C_n}(\theta_1) - \text{prox}_{h_n}^{C_n}(\theta_2)\|_{C_n}.$$

Therefore,  $\text{prox}_{h_n}^{C_n}$  is firmly nonexpansive in the  $C_n$  norm. Denote by  $\lambda_{max}(C_n)$  and  $\lambda_{min}(C_n)$  the maximum and minimum eigenvalues of  $C_n$ . We have

$$\sqrt{\frac{\lambda_{max}(C_n)}{\lambda_{min}(C_n)}} \|\theta_1 - \theta_2\| \geq \|\text{prox}_{h_n}^{C_n}(\theta_1) - \text{prox}_{h_n}^{C_n}(\theta_2)\|.$$

Because  $C_n^{-1} H \rightarrow^P I$  and  $H > 0$ ,  $\sqrt{\lambda_{max}(C_n)/\lambda_{min}(C_n)}$  converges in probability to some constant. It follows that there exists a  $K$  such that for any  $\epsilon > 0$ ,  $n$  large enough implies that  $\sqrt{\lambda_{max}(C_n)/\lambda_{min}(C_n)} < K$  with probability  $1 - \epsilon$ . We conclude that with high probability  $\text{prox}_{h_n}^{C_n}$  is  $K$ -Lipschitz continuous.

We next want to show  $\sqrt{n} \|\hat{\theta}_{ose} - \hat{\theta}\| = o_P(1)$ . From the definition of  $\hat{\theta}_{ose}$ , we have

$$\begin{aligned} \sqrt{n} \|\hat{\theta}_{ose} - \hat{\theta}\| &\leq \sqrt{n} \left\| \text{prox}_{h_n}^{C_n}(\hat{\theta}_{init} - C_n^{-1} \nabla g_n(\hat{\theta}_{init})) - \text{prox}_{h_n}^{C_n}(\hat{\theta} - C_n^{-1} \nabla g_n(\hat{\theta})) \right\| \\ &\quad + \sqrt{n} \left\| \text{prox}_{h_n}^{C_n}(\hat{\theta} - C_n^{-1} \nabla g_n(\hat{\theta})) - \hat{\theta} \right\|. \end{aligned}$$

Downloaded from informas.org by [205.155.65.226] on 09 October 2025, at 14:42. For personal use only, all rights reserved.

Recall from Section 1.2 that  $\widehat{\theta}$  is a fixed point of scaled proximal gradient descent. Therefore, the final term in the sum is  $o_P(1)$ , and we can simplify as follows:

$$= \sqrt{n} \left\| \text{prox}_{h_n}^{C_n} \left( \widehat{\theta}_{init} - C_n^{-1} \nabla g_n(\widehat{\theta}_{init}) \right) - \text{prox}_{h_n}^{C_n} \left( \widehat{\theta} - C_n^{-1} \nabla g_n(\widehat{\theta}) \right) \right\| + o_P(1).$$

From the  $\sqrt{n}$ -consistency of  $\widehat{\theta}_{init}$  and the aforementioned Lipschitz continuity result, we can restrict our considerations to the event where  $\sqrt{n} \|\widehat{\theta}_{init} - \widehat{\theta}\| < M$  for some  $M$  and  $\text{prox}_{h_n}^{C_n}$  is Lipschitz continuous with constant  $K$ , because its complement can be made to have arbitrarily small probability. On this event, assumption (i) holds and the previous display becomes the following:

$$\begin{aligned} &\leq K \sqrt{n} \left\| \widehat{\theta}_{init} - C_n^{-1} \nabla g_n(\widehat{\theta}_{init}) - \left( \widehat{\theta} - C_n^{-1} \nabla g_n(\widehat{\theta}) \right) \right\| + o_P(1) \\ &= K \sqrt{n} \left\| \left( I - C_n^{-1} H \right) \left( \widehat{\theta}_{init} - \widehat{\theta} \right) + C_n^{-1} H \left( \widehat{\theta}_{init} - \widehat{\theta} \right) - C_n^{-1} \left( \nabla g_n(\widehat{\theta}_{init}) - \nabla g_n(\widehat{\theta}) \right) \right\| + o_P(1) \\ &\leq K \left\| I - C_n^{-1} H \right\| \cdot \sqrt{n} \left\| \widehat{\theta}_{init} - \widehat{\theta} \right\| + K \left\| C_n^{-1} \right\| \cdot \sqrt{n} \left\| \nabla g_n(\widehat{\theta}_{init}) - \nabla g_n(\widehat{\theta}) - H \left( \widehat{\theta}_{init} - \widehat{\theta} \right) \right\| + o_P(1). \end{aligned} \quad (3)$$

From assumption (ii),  $\left\| I - C_n^{-1} H \right\| \rightarrow^P 0$ . Because  $\widehat{\theta}_{init}$  is  $\sqrt{n}$ -consistent for  $\widehat{\theta}$ ,  $\sqrt{n} \left\| \widehat{\theta}_{init} - \widehat{\theta} \right\| = O_P(1)$ . Further,  $K \left\| C_n^{-1} \right\| = O_P(1)$  because of assumption (ii). Lastly, assumption (i) gives that

$$\sqrt{n} \left\| \nabla g_n(\widehat{\theta}_{init}) - \nabla g_n(\widehat{\theta}) - H \left( \widehat{\theta}_{init} - \widehat{\theta} \right) \right\| = o_P(1).$$

Therefore the expression in (3) is  $o_P(1)$  by Slutsky's theorem. This concludes the proof.  $\square$

We comment briefly on the assumptions in Theorem 3. The convexity of  $h$  was only used to establish the Lipschitz continuity of the scaled proximal operator. By restricting to the event  $\{\sqrt{n} \|\widehat{\theta}_{init} - \widehat{\theta}\|\}$ , it suffices to allow  $h$  to belong to a wider class of functions, such as lower-semicontinuous and prox-regular, where it can be established that the scaled proximal operator is locally Lipschitz continuous (Hare and Sagastizábal [19, theorem 1]).

Assumption (i) is analogous to a similar condition in Van der Vaart [42], where it is used to demonstrate asymptotic optimality of maximum likelihood estimators and one-step estimators. Despite its resemblance to a condition guaranteeing that  $\nabla g_n$  is differentiable with Jacobian  $H$ , the probabilistic nature of the limit allows it to be satisfied even when  $\nabla g_n$  is not differentiable. In this case, applying the well-known *discretization trick* gives that differentiability in quadratic mean and implies assumption (i) when  $\theta$  converges in probability to some nonrandom  $\theta_0$  (see Le Cam [26] for details).

Assumption (ii) is slightly stronger than necessary; we only require that

$$\sqrt{n} \left( I - C_n^{-1} H \right) \left( \widehat{\theta}_{init} - \widehat{\theta} \right) \rightarrow^P 0.$$

This assumption has an interesting relationship with the Dennis-Moré criterion, a condition for the approximation of Hessians in deterministic variable metric methods (Dennis and Moré [13]). The Dennis-Moré criterion states that the approximation  $C_n$  of a Hessian  $H$  satisfies

$$\frac{\left\| (C_n - H) \left( \widehat{\theta}_{k+1} - \widehat{\theta}_k \right) \right\|}{\left\| \widehat{\theta}_{k+1} - \widehat{\theta}_k \right\|} \rightarrow 0,$$

where  $\widehat{\theta}_{k+1}$  and  $\widehat{\theta}_k$  are iterates of a variable metric method. Assumption (ii) is implied by a probabilistic version of the Dennis-Moré criterion, where iterates are replaced by the  $\widehat{\theta}_{init}$  and  $\widehat{\theta}$ . Indeed, if

$$\frac{\left\| (C_n - H) \left( \widehat{\theta}_{init} - \widehat{\theta} \right) \right\|}{\left\| \widehat{\theta}_{init} - \widehat{\theta} \right\|} = o_P(1),$$

then using the  $\sqrt{n}$ -consistency of  $\widehat{\theta}_{init}$ ,

$$\sqrt{n} \left\| (C_n - H) \left( \widehat{\theta}_{init} - \widehat{\theta} \right) \right\| = O \left( \frac{\left\| (C_n - H) \left( \widehat{\theta}_{init} - \widehat{\theta} \right) \right\|}{\left\| \widehat{\theta}_{init} - \widehat{\theta} \right\|} \right) = o_P(1),$$

yielding assumption (ii) whenever  $\|C_n\|$  is bounded in probability.

We also note that  $\widehat{\theta}$  need not be a minimizer of  $g_n + h_n$  for the conclusion of Theorem 3 to hold. Indeed, the only required property of  $\widehat{\theta}$  is that it is approximately a fixed point of scaled proximal gradient descent. That is,

$$\sqrt{n} \left\| \text{prox}_{h_n}^{C_n} \left( \widehat{\theta} - C_n^{-1} \nabla g_n(\widehat{\theta}) \right) - \widehat{\theta} \right\| = o_P(1). \quad (4)$$

This flexibility allows Theorem 3 to generalize to stationary points, because  $g_n$  need not be convex. We formalize this result in the following corollary.

**Corollary 1.** *In the setting of Theorem 3, if  $\widehat{\theta}$  satisfies (4) as an approximate fixed point of proximal gradient descent, then the conclusion of Theorem 3 holds (even if  $\widehat{\theta}$  is not necessarily a minimizer of  $g_n + h_n$ ).*

We conclude this subsection with a result that permits the application of Theorem 3 to regularized maximum likelihood in the setting of i.i.d. observations from Section 1.1. The following result establishes the asymptotic equivalence of the regularized maximum likelihood estimator and the one-step estimator derived from scaled proximal gradient descent. Its assumptions are classical and verifiable for a large class of parametric models (Van der Vaart [42, section 5.6]).

**Proposition 1.** *Assume that there is some  $\theta_0 \in \Theta$  such that the estimator  $\widehat{\theta}$  in (1) is  $\sqrt{n}$ -consistent for  $\theta_0$ . Assume also that  $X_i \sim^{i.i.d.} p_{\theta^*}$  for some  $\theta^* \in \Theta$ . Let  $g_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i)$ , and let  $h_n$  be proper and convex  $P$ -almost surely. Assume that*

- (i)  $\log p_{\theta}$  is three-times differentiable for every  $\theta$ ;
- (ii) the matrix  $E[\nabla^2 \log p_{\theta_0}]$  exists and is negative definite (so that the Fisher Information  $I_{\theta_0}$  exists and is nondegenerate);
- (iii) the third-order partial derivatives of  $\log p_{\theta}$  are dominated by a fixed integrable function for every  $\theta$  in a neighborhood of  $\theta_0$ ;
- (iv)  $C_n > 0$  and  $-C_n^{-1} E[\nabla^2 \log p_{\theta_0}] \rightarrow^P I$ .

Then for any  $\sqrt{n}$ -consistent estimator,  $\widehat{\theta}_{init}$  of  $\widehat{\theta}$ ,  $\widehat{\theta}$  and the one-step estimator  $\widehat{\theta}_{ose}$  of Equation (2) are asymptotically equivalent.

**Proof.** The result follows from applying Theorem 3 under these assumptions. Thus, we need to show that for each  $M > 0$  there is positive definite  $H$  such that

$$\sup_{\sqrt{n} \|\theta - \widehat{\theta}\| < M} \sqrt{n} (\nabla g_n(\theta) - \nabla g_n(\widehat{\theta})) - \sqrt{n} H(\theta - \widehat{\theta}) \rightarrow^P 0.$$

Take  $H = -E[\nabla^2 \log p_{\theta_0}]$ , so that assumption (iv) is equivalent to Theorem 3's assumption (ii). Fix  $M > 0$ . By the  $\sqrt{n}$ -consistency of  $\widehat{\theta}$ ,  $\sqrt{n} \|\widehat{\theta} - \theta_0\| < K$  for some  $K$ , because its complement can be made to have arbitrarily small probability. We thus have  $\sqrt{n} \|\theta - \theta_0\| < M + K$  when  $\sqrt{n} \|\theta - \widehat{\theta}\| < M$ . We can expand with the triangle inequality to give

$$\begin{aligned} & \left\| \sup_{\sqrt{n} \|\theta - \widehat{\theta}\| < M} \sqrt{n} (\nabla g_n(\theta) - \nabla g_n(\widehat{\theta})) - \sqrt{n} H(\theta - \widehat{\theta}) \right\| \\ & \leq \sup_{\sqrt{n} \|\theta - \widehat{\theta}\| < M} \left\{ \left\| \sqrt{n} (\nabla g_n(\theta) - \nabla g_n(\theta_0)) - \sqrt{n} \nabla^2 g_n(\theta_0) (\theta - \theta_0) \right\| \right. \\ & \quad + \left\| \sqrt{n} (\nabla g_n(\theta_0) - \nabla g_n(\widehat{\theta})) - \sqrt{n} \nabla^2 g_n(\theta_0) (\theta_0 - \widehat{\theta}) \right\| \\ & \quad + \left\| \sqrt{n} (\nabla^2 g_n(\theta_0) - H) (\theta - \theta_0) \right\| + \left\| \sqrt{n} (\nabla^2 g_n(\theta_0) - H) (\theta_0 - \widehat{\theta}) \right\| \left. \right\} \\ & \leq \sup_{\sqrt{n} \|\theta - \theta_0\| < M+K} \left\| \sqrt{n} (\nabla g_n(\theta) - \nabla g_n(\theta_0)) - \sqrt{n} \nabla^2 g_n(\theta_0) (\theta - \theta_0) \right\| \\ & \quad + \left\| \sqrt{n} (\nabla g_n(\theta_0) - \nabla g_n(\widehat{\theta})) - \sqrt{n} \nabla^2 g_n(\theta_0) (\theta_0 - \widehat{\theta}) \right\| \\ & \quad + \sup_{\sqrt{n} \|\theta - \theta_0\| < M+K} \left\| \sqrt{n} (\nabla^2 g_n(\theta_0) - H) (\theta - \theta_0) \right\| \\ & \quad + \left\| \sqrt{n} (\nabla^2 g_n(\theta_0) - H) (\theta_0 - \widehat{\theta}) \right\|. \end{aligned} \tag{5}$$

We have  $\nabla^2 g_n(\theta_0) = -\frac{1}{n} \sum_{i=1}^n \nabla^2 \log p_{\theta_0}(X_i)$ , which converges almost surely to  $H$  by the law of large numbers. Because both  $\sqrt{n}(\theta - \theta_0)$  and  $\sqrt{n}(\widehat{\theta} - \theta_0)$  are bounded in probability, the third and fourth expressions in (5) are  $o_P(1)$ .

We next focus on the first term in the (5). We have

$$\begin{aligned} & \sqrt{n}(\nabla g_n(\theta) - \nabla g_n(\theta_0)) - \sqrt{n}\nabla^2 g_n(\theta_0)(\theta - \theta_0) \\ &= -\frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla \log p_\theta(X_i) - \nabla \log p_{\theta_0}(X_i) - \nabla^2 \log p_{\theta_0}(X_i)(\theta - \theta_0). \end{aligned}$$

By Taylor’s theorem and assumption (iii),

$$\left\| \nabla \log p_\theta(X_i) - \nabla \log p_{\theta_0}(X_i) - \nabla^2 \log p_{\theta_0}(X_i)(\theta - \theta_0) \right\| \leq C(X_i) \|\theta - \theta_0\|^2$$

for some integrable function  $C$ . Therefore,

$$\begin{aligned} & \left\| \sqrt{n}(\nabla g_n(\theta) - \nabla g_n(\theta_0)) - \sqrt{n}\nabla^2 g_n(\theta_0)(\theta - \theta_0) \right\| \\ & \leq n^{-1/2} \|\theta - \theta_0\|^2 \sum_{i=1}^n C(X_i) \\ & = n \|\theta - \theta_0\|^2 n^{-3/2} \sum_{i=1}^n C(X_i). \end{aligned}$$

Because  $n \|\theta - \theta_0\|^2$  is bounded in the supremum and  $\frac{1}{n} \sum_{i=1}^n C(X_i)$  is bounded in probability by the law of large numbers, we conclude that

$$\sup_{\sqrt{n}\|\theta - \theta_0\| < M+K} \left\| \sqrt{n}(\nabla g_n(\theta) - \nabla g_n(\theta_0)) - \sqrt{n}\nabla^2 g_n(\theta_0)(\theta - \theta_0) \right\| \rightarrow^P 0.$$

The term

$$\left\| \sqrt{n}(\nabla g_n(\theta_0) - \nabla g_n(\hat{\theta})) - \sqrt{n}\nabla^2 g_n(\theta_0)(\theta_0 - \hat{\theta}) \right\|$$

is  $o_P(1)$  by similar reasoning, where we use the  $\sqrt{n}$ -consistency of  $\hat{\theta}$  instead of the  $\sqrt{n}\|\theta - \theta_0\| < M + K$  condition in the supremum. We conclude that

$$\left\| \sup_{\sqrt{n}\|\theta - \hat{\theta}\| < M} \left( \sqrt{n}(\nabla g_n(\theta) - \nabla g_n(\hat{\theta})) - \sqrt{n}H(\theta - \hat{\theta}) \right) \right\| = o_P(1),$$

which completes the proof.  $\square$

In Proposition 1, we can take  $\hat{\theta}_{init}$  to be any  $\sqrt{n}$ -consistent estimator of  $\theta_0$  (which makes it  $\sqrt{n}$ -consistent for  $\hat{\theta}$ ), such as a moment estimator. We emphasize that Proposition 1 permits  $\theta_0 \neq \theta^*$ , which is of interest in structured inference problems where having estimates with certain properties (i.e., sparsity) may be more important than converging to the truth.

## 2.1. Stopping Condition

The following result provides a recipe for finding a  $\sqrt{n}$ -consistent estimator  $\hat{\theta}_{init}$ . Given mild assumptions on the composite function  $f_n = g_n + h_n$ ,  $\hat{\theta}_{init}$  can be generated from a number of scaled proximal gradient steps and an easily verifiable stopping condition. Its proof is in Appendix A.

**Proposition 2.** *Let  $f_n = g_n + h_n$  be a composite function, and let  $\hat{\theta}$  be a minimizer of  $f_n$ . Assume that  $h_n$  is proper and convex  $P$ -almost surely. Assume further that, with high probability, the function  $g_n$  is strongly convex with parameter  $m$  and  $\nabla g_n$  is  $M$ -Lipschitz continuous. Let  $C_n > 0$  be the matrix used in the scaled proximal gradient step. Assume that there is a constant  $L$  such that  $C_n \leq LI$  with high probability. Then for any sequence  $r_n$  such that  $\|\hat{\theta}_{ose} - \hat{\theta}_{init}\| = O_P(r_n)$ , we have  $\|\hat{\theta}_{init} - \hat{\theta}\| = O_P(r_n)$ .*

Proposition 2 allows us to generate  $\sqrt{n}$ -consistent estimators from iterations of scaled proximal gradient descent. In practice, one can use an off-the-shelf implementation of scaled proximal gradient descent, such as the popular R package `glmnet` for penalized generalized linear models (Friedman et al. [17]), and take  $\hat{\theta}_{init}$  to be any point for which the step length is less than  $c/\sqrt{n}$  for some chosen constant  $c$ . We also note that Proposition 2 applies when with high probability  $\hat{\theta}, \hat{\theta}_{init}, \hat{\theta}_{ose}$  all lie in some set where the Lipschitz and strong convexity conditions hold.

### 3. Proximal Descent

In this section, we consider one-step estimators formed by minimizing an objective with scaled proximal descent. In contrast to the previous section, the objective function is written as a single function instead of the sum in the composite model. Consider the estimator

$$\widehat{\theta} = \arg \min_{\theta} f_n(\theta), \quad (6)$$

where, as in Section 2,  $f_n : \Omega \times \Theta \rightarrow \mathbb{R} \cup \{\infty\}$  is a sequence of random functions on  $\Theta \subseteq \mathbb{R}^d$ , defined on a probability space  $(\Omega, \mathcal{A}, P)$ . For a sequence of random  $d \times d$  scaling matrices  $C_n$  and an initial estimator  $\widehat{\theta}_{init}$ , we define a one-step estimator from proximal descent as

$$\widehat{\theta}_{ose} = \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}). \quad (7)$$

Compared with the scaled proximal gradient descent method in the previous section, this one-step estimator is a scaled proximal descent method (i.e., it omits the gradient step). We next provide a result that shows that  $\widehat{\theta}_{ose}$  in (7) is asymptotically equivalent to  $\widehat{\theta}$  from (6). This result is an analogue of Theorem 3 for scaled proximal descent.

**Theorem 4.** Consider the estimator  $\widehat{\theta}$  in Equation (6) and  $\widehat{\theta}_{ose}$  in Equation (7), where  $f_n$  is such that both minimizers exist almost surely. Assume the following:

- (i) We have  $\lambda_{max}(C_n) = o_P(1)$ .
- (ii) There is a constant  $L$  such that for each constant  $M$ , the function  $\text{prox}_{f_n}^{C_n}$  is  $L$ -Lipschitz continuous on  $\{\theta : \|\theta - \widehat{\theta}\| \leq M/\sqrt{n}\}$  with high probability.
- (iii) There is a constant  $m$  such that for each constant  $M$ , the function  $f_n$  is strongly convex with modulus  $m$  on  $\{\theta : \|\theta - \widehat{\theta}\| \leq M/\sqrt{n}\}$  with high probability.

Then for any  $\sqrt{n}$ -consistent estimator  $\widehat{\theta}_{init}$  of  $\widehat{\theta}$ ,  $\widehat{\theta}$  and  $\widehat{\theta}_{ose}$  are asymptotically equivalent.

**Proof.** As a global minimizer of  $f_n$ ,  $\widehat{\theta}$  is a fixed point of  $\text{prox}_{f_n}^{C_n}$ . Assume that  $\sqrt{n}\|\widehat{\theta} - \widehat{\theta}_{init}\| < M$  for some constant  $M$ , because its complement has arbitrarily small probability by the  $\sqrt{n}$ -consistency of  $\widehat{\theta}_{init}$ .

We need to show that  $\sqrt{n}\|\widehat{\theta}_{ose} - \widehat{\theta}\| = o_P(1)$ . From the definition of  $\widehat{\theta}_{ose}$ , we have the following:

$$\begin{aligned} \sqrt{n}\|\widehat{\theta}_{ose} - \widehat{\theta}\| &\leq \sqrt{n}\left\|\widehat{\theta}_{ose} - \text{prox}_{f_n}^{C_n}(\widehat{\theta})\right\| + \sqrt{n}\left\|\text{prox}_{f_n}^{C_n}(\widehat{\theta}) - \widehat{\theta}\right\| \\ &= \sqrt{n}\left\|\text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta})\right\| + \sqrt{n}\left\|\text{prox}_{f_n}^{C_n}(\widehat{\theta}) - \widehat{\theta}\right\| \\ &= \sqrt{n}\left\|\text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta})\right\| + o_P(1). \end{aligned} \quad (8)$$

We will focus our remaining efforts on the first term in (8). Assume that  $\sqrt{n}\|\widehat{\theta}_{init} - \widehat{\theta}\| < M$  for some  $M$ , because its complement can be made to have negligible probability. By assumption (ii) we can further assume that  $\text{prox}_{f_n}^{C_n}$  is single-valued. Finally, Lipschitz continuity of the scaled prox and assumption (iii) allow us to assume that  $f_n$  is strongly convex with modulus  $m$  on some set containing both  $\text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init})$  and  $\text{prox}_{f_n}^{C_n}(\widehat{\theta})$ . From the definition of the proximal operator, we have

$$\widehat{\theta}_{init} \in \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) + C_n^{-1} \partial f_n \left( \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) \right) \quad \text{and} \quad \widehat{\theta} \in \text{prox}_{f_n}^{C_n}(\widehat{\theta}) + C_n^{-1} \partial f_n \left( \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right). \quad (9)$$

We also have

$$\begin{aligned} &\lambda_{max}(C_n) \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right\| \|\widehat{\theta}_{init} - \widehat{\theta}\| \\ &\geq \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right\|_{C_n} \|\widehat{\theta}_{init} - \widehat{\theta}\|_{C_n}, \end{aligned} \quad (10)$$

so from (9), there are  $u \in \partial f_n(\text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}))$  and  $v \in \partial f_n(\text{prox}_{f_n}^{C_n}(\widehat{\theta}))$  such that (10) is bounded below as

$$\geq \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right\|_{C_n} \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - C_n^{-1}u - \left( \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - C_n^{-1}v \right) \right\|_{C_n}.$$

Applying the Cauchy-Schwarz inequality,

$$\begin{aligned} &\geq \left( \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right)^T C_n \left( \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) + C_n^{-1}(u - v) \right) \\ &= \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right\|_{C_n}^2 + \left( \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right)^T (u - v). \end{aligned}$$

Invoking strong convexity of  $f_n$ , we have that  $\partial f_n$  is strongly monotone for some constant  $m$ , so we continue as follows:

$$\begin{aligned} &\geq \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right\|_{C_n}^2 + m \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right\|_{C_n}^2 \\ &\geq (\lambda_{\min}(C_n) + m) \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right\|_{C_n}^2. \end{aligned}$$

From this chain of inequalities, we conclude that

$$\frac{\lambda_{\max}(C_n)}{\lambda_{\min}(C_n) + m} \|\widehat{\theta}_{init} - \widehat{\theta}\| \geq \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right\|_{C_n}.$$

Therefore, if  $\lambda_{\max}(C_n) \rightarrow^P 0$ , because  $\sqrt{n} \|\widehat{\theta}_{init} - \widehat{\theta}\|$  is bounded in probability, we have

$$\sqrt{n} \left\| \text{prox}_{f_n}^{C_n}(\widehat{\theta}_{init}) - \text{prox}_{f_n}^{C_n}(\widehat{\theta}) \right\|_{C_n} = o_P(1).$$

We conclude that

$$\sqrt{n} \|\widehat{\theta}_{ose} - \widehat{\theta}\| = o_P(1),$$

which completes the proof.  $\square$

Similar to Theorem 3, we note that  $\widehat{\theta}$  need not be a global minimizer of  $f_n$  to apply Theorem 4 and instead satisfy the approximate fixed-point condition  $\text{prox}_{f_n}^{C_n}(\widehat{\theta}) - \widehat{\theta} = o_P(n^{-1/2})$ .

We also include a proposition that permits the application of Theorem 4 in the setting of parametric estimation from i.i.d. observations introduced in Section 1.1.

**Proposition 3.** Let  $X_1, \dots, X_n \sim^{i.i.d.} p_{\theta^*}$  and  $f_n = -\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i)$ . Assume the following:

- (i) We have  $\log p_{\theta} \in C^2$  for each  $X$ . Further, partial derivatives may be passed under the integral in  $\int p_{\theta}(x) dx$ .
- (ii) There exists a function  $K(x)$  such that  $E[K(X_1)] < \infty$  and each component of  $\nabla^2 \log p_{\theta_0}$  is bounded in absolute value by  $K$  uniformly in some neighborhood of  $\theta_0$ .
- (iii) The matrix  $E[\nabla^2 \log p_{\theta_0}]$  exists and is negative definite (so that the Fisher information  $I_{\theta_0}$  exists and is nongenerate). Further, the Fisher information is continuous at  $\theta_0$ .

Then Theorem 4 holds with only the first two assumptions of that theorem. That is, if

- (i')  $\lambda_{\max}(C_n) = o_P(1)$ ;
- (ii') there is a constant  $L$  such that for each constant  $M$ , the function  $\text{prox}_{f_n}^{C_n}$  is  $L$ -Lipschitz continuous on  $\{\theta : \|\theta - \widehat{\theta}\| \leq M/\sqrt{n}\}$  with high probability.

Then for any  $\sqrt{n}$ -consistent estimator  $\widehat{\theta}_{init}$  of  $\widehat{\theta}$ ,  $\widehat{\theta}$  from (6) and  $\widehat{\theta}_{ose}$  from (7) are asymptotically equivalent.

**Proof.** For brevity, we denote  $\log p_\theta$  by  $\ell_\theta$ . We first show that there exists  $m > 0$  such that for all  $\|h\| \leq B$ , it holds with high probability that  $f_n := -\sum_{i=1}^n \ell_{\theta_0+h/\sqrt{n}}(X_i)$  is strongly convex as a function of  $h$  with parameter  $m$ . Take  $m$  to be such that  $\lambda_{\min}(I_{\theta_0}) > m > 0$ .

Assumption (i) gives the well-known result that  $I_\theta = -E[\nabla^2 \ell_\theta(X_1)]$ . By Taylor's theorem,

$$\sum_{i=1}^n \ell_{\theta_0+h/\sqrt{n}}(X_i) = \sum_{i=1}^n \ell_{\theta_0}(X_i) + h^T \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla \ell_{\theta_0}(X_i) \right) + \frac{1}{2} h^T \left( \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell_{\theta_0+t_i h/\sqrt{n}}(X_i) \right) h$$

for some  $t_i \in (0, 1)$ . We note that the  $t_i$  are random because they depend on  $X_i$ .

Our analysis will focus on the quadratic term. We apply the uniform law of large numbers (Ferguson [15, theorem 16a]) and the continuity of  $I_\theta$  to see that

$$\begin{aligned} & \sup_{\|h\| \leq B} \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell_{\theta_0+t_i h/\sqrt{n}}(X_i) - E_{\theta^*} [\nabla^2 \ell_{\theta_0}(X_1)] \right\| \\ & \leq \sup_{\|h\| \leq B} \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell_{\theta_0+h/\sqrt{n}}(X_i) - E_{\theta^*} [\nabla^2 \ell_{\theta_0}(X_1)] \right\| \\ & \leq \sup_{\|h\| \leq B} \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell_{\theta_0+h/\sqrt{n}}(X_i) - E_{\theta^*} [\nabla^2 \ell_{\theta_0+h/\sqrt{n}}(X_1)] \right\| \\ & \quad + \sup_{\|h\| \leq B} \left\| E_{\theta^*} [\nabla^2 \ell_{\theta_0+h/\sqrt{n}}(X_1)] - E_{\theta^*} [\nabla^2 \ell_{\theta_0}(X_1)] \right\| \\ & = \sup_{\|h\| \leq B} \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell_{\theta_0+h/\sqrt{n}}(X_i) + E_{\theta^*} [\nabla^2 \ell_{\theta_0+h/\sqrt{n}}(X_1)] \right\| \\ & \quad + \sup_{\|h\| \leq B} \|I_{\theta_0+h/\sqrt{n}} - I_{\theta_0}\|. \end{aligned}$$

The first inequality in the previous display follows from  $t_i \in (0, 1)$ . Because both of the terms in the last equation converge to zero in probability, the Hessian term in the Taylor expansion of  $\ell_{\theta_0+h/\sqrt{n}}$  converges in probability to  $-I_{\theta_0}$ . Thus, for each  $\epsilon > 0$ , we can choose  $N$  such that  $n \geq N$  implies

$$\frac{1}{n} \sum_{i=1}^n \nabla^2 \ell_{\theta_0+t_i h/\sqrt{n}}(X_i) \leq -I_{\theta_0} + \epsilon I$$

with arbitrarily high probability. Thus,  $f_n$  is strongly convex with high probability for  $m > 0$ .

We have shown that, for each constant  $M$ , the function  $f_n$  is strongly convex with high probability on  $\{\theta : \|\theta - \hat{\theta}\| \leq M/\sqrt{n}\}$ . This is assumption (iii) in Theorem 4, so the result is proved.  $\square$

## 4. The Moreau Envelope as a Statistical Smoother

In this section, we interpret our contributions in the context of smoothing irregularities in a statistical objective. Theorems 3 and 4 both provide equivalences between OSE and MLE estimators but additionally allow an important connection to infimal convolution and the scaled Moreau envelope as a statistical smoother.

For two extended real-valued functions  $f_1 : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  and  $f_2 : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$ , the infimal convolution  $f_1 \# f_2 : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  is

$$(f_1 \# f_2)(x) = \inf_{x_1+x_2=x} \{f_1(x_1) + f_2(x_2)\}.$$

As long as the infimum in the definition of  $(f_1 \# f_2)$  is attained, the epigraph of the inf-convolution is the Minkowski sum of the epigraphs of  $f_1$  and  $f_2$  (Rockafellar and Wets [37]). This interpretation allows the infimal convolution to be used as a smoothing operation, where  $f_1 \# f_2$  gives a smoothed version of  $f_1$  for  $f_2$  chosen with an epigraph satisfying certain regularity properties. See either Burke and Hoheisel [10, 11], or Xu et al. [43] for recent development of inf-convolution as a smoother, Rockafellar and Wets [37] for a more general overview of the properties of inf-convolution, and Beck and Teboulle [6] for more on Moreau envelope as a smoother, as well as its connection to the proximal operator.

The Moreau envelope is a special case of inf-convolution, where a function is inf-convolved against a squared  $\ell_2$  norm. For  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  and positive definite matrix  $C \in \mathbb{R}^{d \times d}$ , the scaled Moreau envelope  $e_{Cf} : \mathbb{R}^d \rightarrow \mathbb{R}$  is

$$e_{Cf}(x) = \inf_{w \in \mathbb{R}^d} \left\{ f(w) + \frac{1}{2} \|x - w\|_C^2 \right\}.$$

The (unscaled) Moreau envelope is the scaled Moreau envelope with  $C = (1/\lambda)I$  for some positive scalar  $\lambda$ . The scaled Moreau envelope is closely linked to the scaled prox, because for any  $w^* \in \text{prox}_f^C(x)$ ,

$$e_{Cf}(x) = f(w^*) + \frac{1}{2} \|x - w^*\|_C^2.$$

Despite the tendency of minimization to destroy smoothness, the Moreau envelope is smooth for convex functions  $f$  (Rockafellar and Wets [37]). In the nonconvex case, smoothness of the Moreau envelope at a point  $x^*$  requires inner continuity<sup>2</sup> of the scaled prox at a point in  $\{x^*\} \times \text{prox}_f^C(x^*)$ . We formalize this result in the following proposition. Its proof can be found in Appendix A.

**Proposition 4** (Differentiability of  $e_{Cf}$ ). *Let  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$  be an extended real-valued function. For a positive definite matrix  $C > 0$ , the Moreau envelope  $e_{Cf}$  is differentiable at  $x^*$  if there exists  $w^* \in \text{prox}_f^C(x^*)$  such that for all sequences  $u^v \rightarrow 0$  there exists  $w^v \in \text{prox}_f^C(x^* + u^v)$  with  $w^v \rightarrow w^*$ . In this case,  $\nabla e_{Cf}(x^*) = C(x^* - w^*)$ .*

It is clear from Proposition 4 that the objective function  $e_{Cf}$  is a smoothed version of the function  $f$  when  $f$  satisfies the required assumptions. Moreover,  $e_{Cf}$  is finite over  $\mathbb{R}^d$  when  $f$  is proper and bounded from below, even though  $f$  itself might take the value  $\infty$ . We also note that  $e_{Cf}$  preserves global minimizers of  $f$ ; if  $x^*$  is a minimizer of  $f$ , then the following chain of inequalities hold for all  $x$  and  $w$ :

$$\begin{aligned} f(x^*) &\leq f(w) + \frac{1}{2} \|x - w\|_C^2; \\ e_{Cf}(x^*) &\leq f(x^*) + \frac{1}{2} \|x^* - x^*\|^2 = f(x^*). \end{aligned}$$

Taking the infimum over  $w$  and  $x$  implies that the sets of minimizers of  $f$  and  $e_{Cf}$  coincide. The smoothing properties of the Moreau envelope are depicted in Figure 1, which also illustrates the coincidence of minimizers.

When the Moreau envelope is differentiable at a point  $x$ , Proposition 4 gives

$$\nabla e_{Cf}(x) = C \left( x - \underset{f}{\text{prox}}(x) \right) \Rightarrow \underset{f}{\text{prox}}(x) = x - C^{-1} \nabla e_{Cf}(x).$$

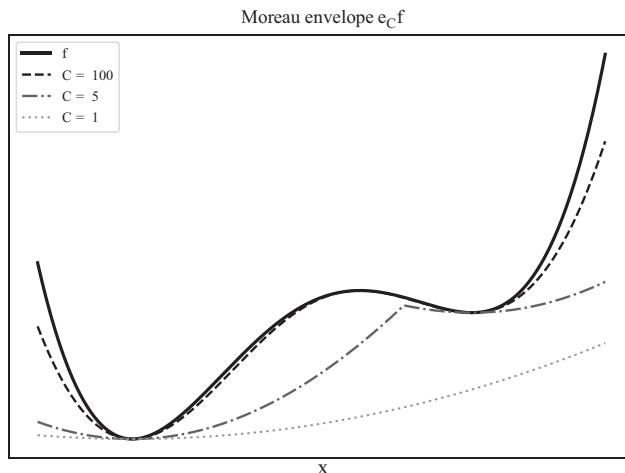
We conclude that the proximal operator is scaled gradient descent applied to the Moreau envelope. This perspective allows us to interpret Theorem 4 in the context of smoothing the likelihood function with the Moreau envelope.

**Corollary 2.** *Under the assumptions of Theorem 4, the one-step estimator*

$$\hat{\theta}_{ose} = \hat{\theta}_{init} - C_n^{-1} \nabla e_{C_n} f_n(\hat{\theta}_{init}),$$

*which is a single iteration of scaled gradient descent applied to the Moreau envelope  $e_{C_n}$ , is asymptotically equivalent to the estimator  $\hat{\theta}$  that minimizes  $f_n$ .*

**Figure 1.** A collection of (unscaled) Moreau envelopes for a nonconvex function.



Interpreted in the context of the Moreau envelope, Corollary 2 provides a recipe for smoothing a negative log-likelihood while retaining the attractive theoretical properties of the maximum likelihood estimator. Thus, the Moreau envelope can be used to smooth a statistical objective, with the goal of removing irrelevant local artifacts while preserving important global structure. Other efforts in this area include the continuation method, also called *graduated optimization*, in which an objective is convolved against a smooth (usually Gaussian) kernel in order to impart additional smoothness in the objective function (Hazan et al. [20], Mobahi et al. [33]). The continuation method also appears independently in the statistics literature as maximum smoothed likelihood estimation (Ionides [22], Eggermont [13]). The main limitation to the practical application of the continuation method is the difficulty associated with computing gradient and curvature information for the smoothed function. Smoothing objectives with the Moreau envelope, on the other hand, has the attractive property that descent steps are evaluations of the scaled proximal operator. Many techniques exist to compute the scaled proximal operator; for examples, see Friedlander and Goh [16], Friedman et al. [17], and Lee et al. [28].

## 5. Experiments and Examples

### 5.1. Estimation with the Cauchy Distribution

In this section, we discuss our results in the context of estimating the location parameter of a Cauchy distributed random variable. We begin by formulating a penalized objective using a MAP (maximum a posteriori) estimator, to which we apply scaled proximal gradient descent. We then alter the problem to remove the prior information and apply scaled proximal descent and Theorem 4 to the MLE problem.

Let  $X_1, \dots, X_n \sim^{i.i.d.} \text{Cauchy}(\theta_0, \sigma_0)$ , where  $\theta_0$  is the location parameter and  $\sigma_0$  is the scale parameter. Assume that  $\sigma_0$  is known and we wish to estimate  $\theta_0$ . Moreover, assume that we incorporate a regularizer through a Laplacian prior on  $\theta_0$ , where  $\sigma_0 \sim \text{Laplace}(0, \gamma)$ . The sample and prior information can be combined using a MAP formulation, where the objective takes the form

$$\min_{\theta} -\frac{1}{n} \sum_{i=1}^n \left\{ \log \left( 1 + \left( \frac{X_i - \theta}{\sigma_0} \right)^2 \right) \right\} + \frac{1}{n\gamma} |\theta|.$$

This problem is nonconvex and has multiple local minima, making its computation through iterative methods difficult. Moreover, the sample mean is not consistent for the location parameter. Despite these challenges, the Cauchy distribution yields a negative log-likelihood  $g_n$  that satisfies the conditions of Proposition 3. Hence,  $\theta_0$  can be estimated using a one-step scaled proximal gradient estimator, where the scaling  $C_n \rightarrow \frac{1}{2\sigma_0}$ , the Fisher information of the Cauchy distribution at  $\theta = 0$ . We will use this example to illustrate the implications of Theorem 3 for nonconvex problems generally. In the remainder, we fix  $\sigma_0 = 20$  and  $\gamma = 1,000$ .

Recall that iterations of scaled proximal gradient descent can also be written as

$$\theta_{k+1} = \arg \min_{\theta} \left\{ g_n(\theta_k) + \nabla g_n(\theta_k)^T (\theta - \theta_k) + h_n(\theta) + \left\| \theta - \theta_k \right\|_{C_n}^2 \right\}.$$

Hence, iterates can be viewed as solving a penalized version of the original problem, where  $g_n$  is linearized around the current iterate. Because this example is univariate, scaled proximal gradient descent is equivalent to the unscaled version. Figure 2 gives a sequence of iterates of scaled proximal gradient descent, as the scaling increases from  $1/400$  to  $1/2\sigma_0$  and the sample size  $n$  increases from 100 to 1,000.

If we omit the Laplacian prior, then we have a maximum likelihood objective without regularization. In this case, the smoothing discussion in the previous section applies, so that scaled proximal descent applied to this objective is scaled gradient descent applied to the Moreau envelope of the negative log-likelihood. Interpreted in this context, Theorem 4 provides theoretical justification for smoothing the Cauchy distribution's negative log-likelihood with the Moreau envelope. Figure 3 illustrates this smoothing for a certain sample.

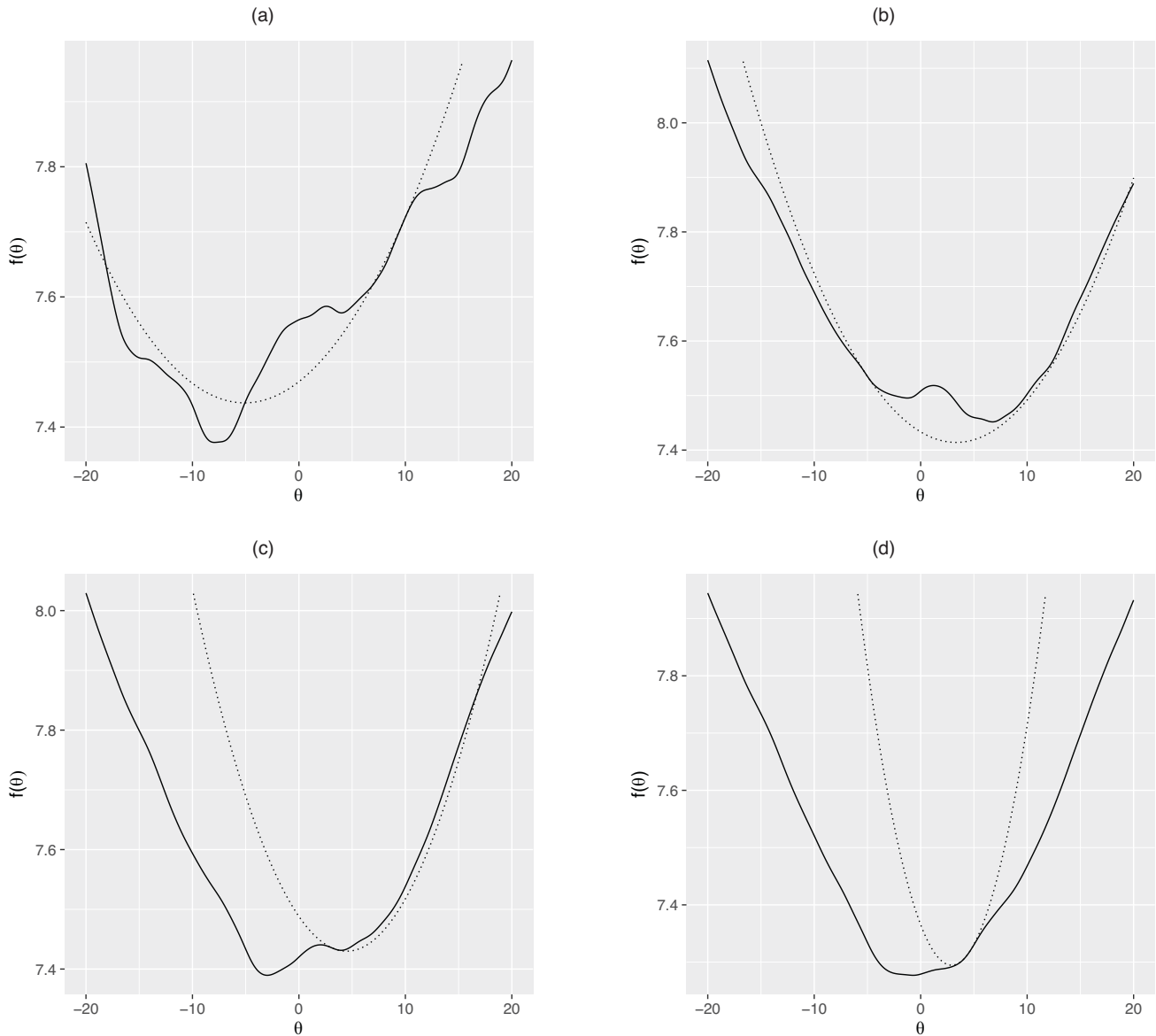
### 5.2. First-Order Methods Are Not Enough

The asymptotic equivalence of the scaled proximal gradient one-step estimator and the MLE prompts the following question: can a one-step estimator without the scaling, such as gradient descent, also be shown to be asymptotically equivalent to the MLE? In this section, we will provide a counterexample demonstrating that, without curvature information, the gradient one-step estimator cannot overcome bias introduced in the distribution of  $\hat{\theta}_{init}$ .

Assume that we want to estimate the mean of a bivariate normal distribution with known variance. Fix  $\mu \in \mathbb{R}^2$ , and let  $X_1, \dots, X_n \sim^{i.i.d.} N(\mu, \Sigma)$ , where

$$\Sigma = \text{diag} \left( (\sigma_1^2, \sigma_2^2)^T \right).$$

**Figure 2.** A sequence of iterates of proximal gradient descent for increasing  $n$  and  $C_n \rightarrow 1/\sigma_0 = 80$ . The solid line gives the objective function, which changes as  $n$  increases. The dashed line gives the function that is minimized to determine the next iterate. Choosing  $C_n$  small in early iterates has a smoothing effect on the sequence.



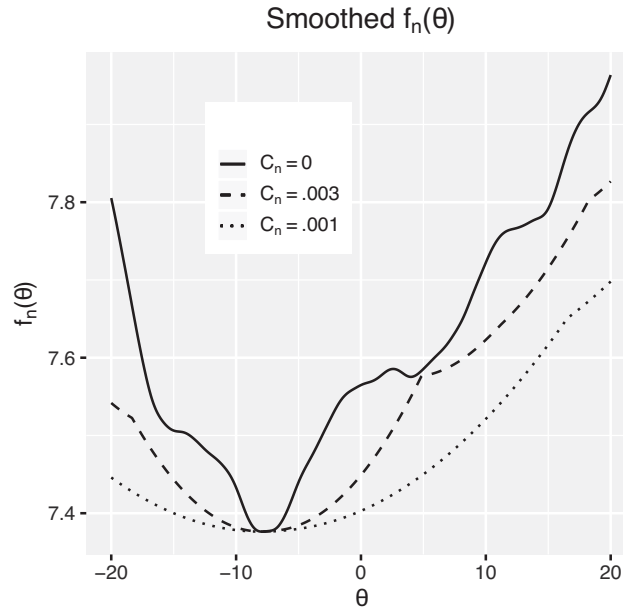
Assume without loss of generality that  $\sigma_1^2 > \sigma_2^2$ . We denote by  $\hat{\theta}_{init}$  the initial estimate and  $\hat{\theta}$  the maximum likelihood estimator of the model. The negative log-likelihood function is quadratic with Hessian  $\Sigma^{-1}$ . Let  $\alpha$  be the step length of gradient descent, which will be specified momentarily. Choose  $\hat{\theta}_{init} \sim U((\mu_1 - \frac{1}{\sqrt{n}}, \mu_1) \times (\mu_2 - \frac{1}{\sqrt{n}}, \mu_2))$  independently of  $X_1, \dots, X_n$ . We have, up to a constant factor

$$f_n(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (\theta - X_i)^\top \Sigma^{-1} (\theta - X_i).$$

Hence,

$$\hat{\theta} = \hat{\theta}_{init} - \alpha \nabla f_n(\hat{\theta}_{init}) = \hat{\theta}_{init} - \alpha \sigma_1^{-2} (\hat{\theta}_{init} - \bar{X}).$$

**Figure 3.** Negative log-likelihood and Moreau envelopes for the location parameter in a Cauchy distributed random variable. In this example,  $\theta_0 = 0$ ,  $\sigma_0 = 20$ , and  $n = 100$ .



Consider  $P(\hat{\theta}_1 > \mu_1)$ , the probability that  $\hat{\theta}$  exceeds  $\mu$  in its first coordinate (where we denote the first coordinate with a subscript):

$$\begin{aligned} P(\hat{\theta}_1 > \mu_1) &= P(\hat{\theta}_{init,1} - \alpha \sum^{-1}(\hat{\theta}_{init,1} - \bar{X}_1) > \mu_1) \\ &= P((1 - \alpha\sigma_1^{-2})(\hat{\theta}_{init,1} - \mu_1) + \alpha\sigma_1^{-2}(\bar{X}_1 - \mu_1) > 0) \\ &= P\left(\bar{X}_1 - \mu_1 > \frac{(1 - \alpha\sigma_1^{-2})(\mu_1 - \hat{\theta}_{init,1})}{\alpha\sigma_1^{-2}}\right). \end{aligned}$$

We have  $\sqrt{n}(\bar{X}_1 - \mu_1) = \sigma_1 Z$ , where  $Z$  is a standard normal. Further,  $\sqrt{n}(\mu_1 - \hat{\theta}_{init,1}) = U$ , where  $U$  is uniform on  $[0, 1]$ . We multiply both sides of the preceding inequality by  $\sqrt{n}$  and rewrite as follows:

$$\begin{aligned} &= P\left(Z > \frac{(1 - \alpha\sigma_1^{-2})U}{\alpha\sigma_1^{-1}}\right) \\ &= P\left(Z > \left(\frac{\sigma_1}{\alpha} - \sigma_1^{-1}\right)U\right). \end{aligned}$$

Note that if we can make  $\sigma_1/\alpha - \sigma_1^{-1} = M$  for some fixed large  $M$ , then integration of the previous display shows that

$$P(\hat{\theta}_1 > \mu_1) = 1 - \Phi(M) + (1 - \exp(-M^2/2))/\sqrt{2\pi}/M.$$

We remark that this probability is independent of  $n$  and can be made arbitrarily small for  $M$  chosen large.

We will consider two common choices of step length  $\alpha$ , fixed step length, and exact line search. The objective function  $f$  has Lipschitz constant  $\sigma_2^{-2}$ . According to Luenberger and Ye [30], the optimal fixed step length in gradient descent is the reciprocal to the objective’s Lipschitz constant, that is,  $\sigma_2^2$  in our example. Therefore,

$$\frac{\sigma_1}{\alpha} - \sigma_1^{-1} = \frac{\sigma_1}{\sigma_2^2} - \sigma_1^{-1}.$$

By fixing  $\sigma_1$  and choosing  $\sigma_2$  small, this expression can be made large, so that the probability  $P(\hat{\theta}_1 > \mu_1)$  is arbitrarily large independent of  $n$ . Because the maximum likelihood estimator  $\bar{X}$  has  $P(\bar{X}_1 > \mu_1)$ ,  $\hat{\theta}_1$  is not asymptotically equivalent to the MLE.

Moreover, we can extend this result beyond the fixed step length to the setting of *exact* step lengths. With a quadratic objective, the exact step length for steepest descent is (see Luenberger and Ye [30])

$$\alpha := \frac{\nabla f_n(\theta)^\top \nabla f_n(\theta)}{\nabla f_n(\theta)^\top \Sigma^{-1} \nabla f_n(\theta)}.$$

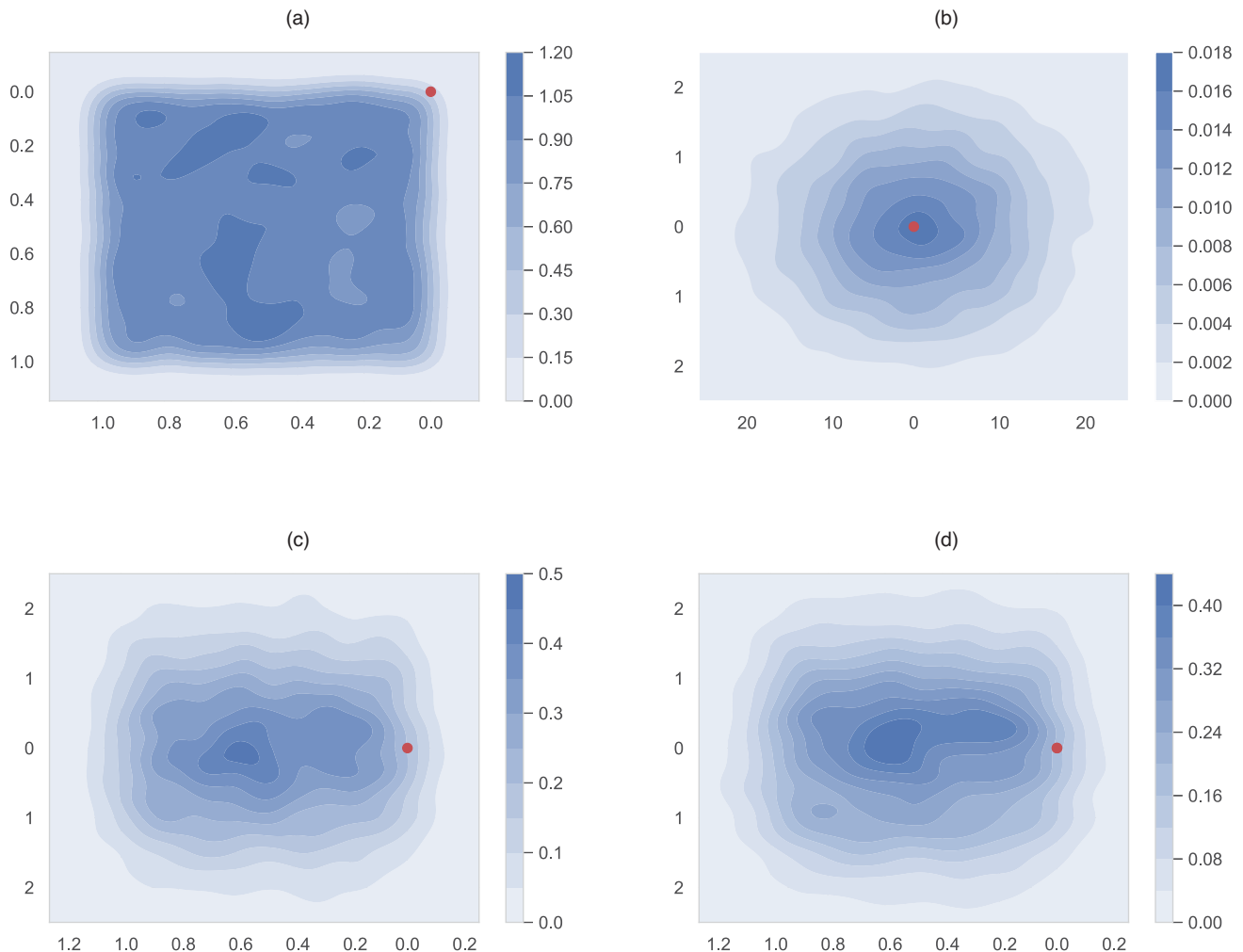
In the aforementioned example, this is

$$\begin{aligned} \alpha &= \frac{(\bar{X} - \theta)^\top \Sigma^{-2} (\bar{X} - \theta)}{(\bar{X} - \theta)^\top \Sigma^{-3} (\bar{X} - \theta)} \\ &= \frac{\sigma_1^{-2} (\bar{X}_1 - \theta_1)^2}{\sigma_1^{-3} (\bar{X}_1 - \theta_1)^2} + \sigma_2^{-3} (\bar{X}_2 - \theta_2)^2 + \frac{\sigma_2^{-2} (\bar{X} - \theta_2)^2}{\sigma_1^{-3} (\bar{X}_1 - \theta_1)^2} + \sigma_2^{-3} (\bar{X}_2 - \theta_2)^2 \\ &\leq \sigma_1 + \sigma_2. \end{aligned}$$

Therefore,

$$\frac{\sigma_1}{\alpha} - \sigma_1^{-1} \geq \frac{\sigma_1}{\sigma_1 + \sigma_2} - \sigma_1^{-1},$$

**Figure 4.** (Color online) Kernel density estimates of the  $\sqrt{n}$ -normalized asymptotic distributions for (a) the starting point, (b) maximum likelihood estimator, (c) the one-step gradient descent estimator with optimal *fixed* step length, and (d) the one-step gradient descent estimator with optimal step length. In each plot, the population mean of  $X$  is given in red. Ten thousand samples were used to construct each density estimate. The estimators in (b)–(d) were each constructed from samples of  $X$  of size 10,000. In this example,  $\sigma_1 = 10$  and  $\sigma_2 = 1$ .



so when  $\sigma_1/(\sigma_1 + \sigma_2) - \sigma_1^{-1} > 0$ ,

$$P(\hat{\theta}_1 > \mu_1) = P\left(Z > \left(\frac{\sigma_1}{\alpha} - \sigma_1^{-1}\right)U\right) \leq P\left(Z > \left(\frac{\sigma_1}{\sigma_1 + \sigma_2} - \sigma_1^{-1}\right)U\right).$$

For  $M$  arbitrarily large, we can choose  $\sigma_1 = 1$  and  $\sigma_2$  arbitrarily small so that  $\frac{\sigma_1}{\sigma_1 + \sigma_2} - \sigma_1^{-1} = M$ . Therefore, we can again make  $P(\hat{\theta}_1 > \mu_1)$  arbitrarily small independently of  $n$ , which shows that again the one-step estimator  $\hat{\theta}_n$  is not asymptotically equivalent to the MLE. Figure 4 gives the empirical distribution of the one-step gradient descent estimator for a number of samples  $X$  and starting points  $\hat{\theta}_{init}$  and for both the optimal fixed and optimal step lengths.

### 5.3. Low-Rank Logistic Regression

We next consider an application in low-rank logistic regression, where we will apply our one-step estimation results for the composite model from Section 2. By using one-step estimation along with the stopping condition presented in Proposition 2, we construct an estimator that has the same large sample performance as the nuclear-norm penalized MLE but which is achieved in a finite number of iterations of scaled proximal gradient descent.

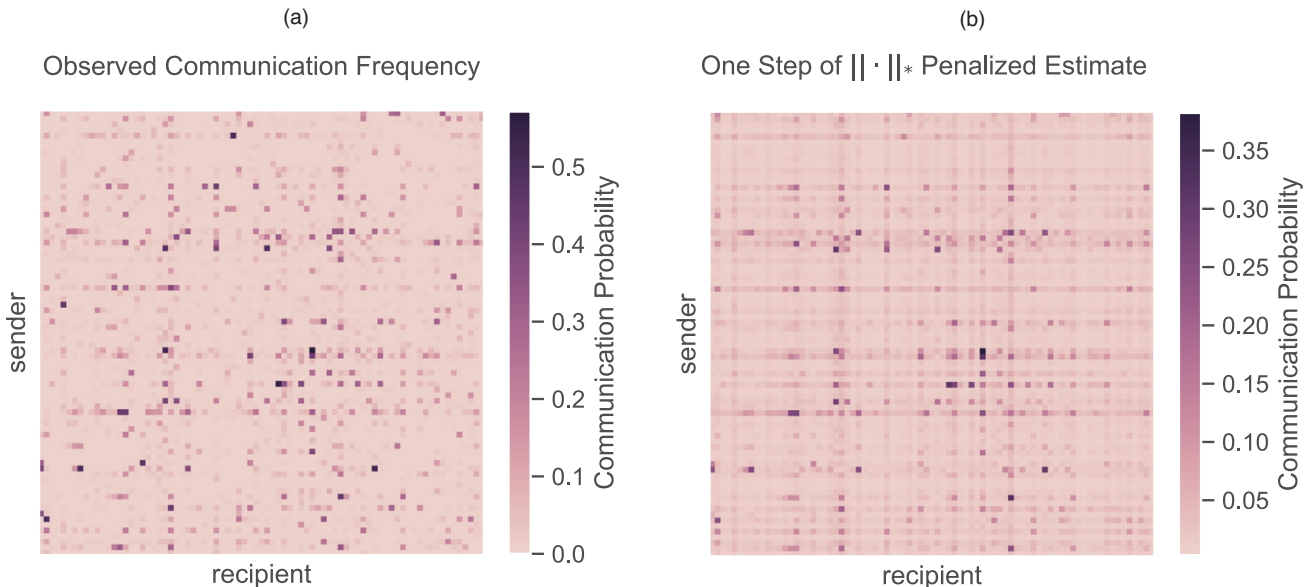
Consider the email-Eu-core data set (Leskovec et al. [29]), formed by a collection of the time indices for emails sent within an academic department. We discretize the time into 49 segments, and we assume that the probability of an email being sent from individual  $i$  to individual  $j$  is constant throughout time and equal to  $P_{ij}$ . Our goal is to estimate  $P$ , the  $N \times N$  matrix of communication probabilities, where  $N$  is the number of individuals in the department. We further assume that the logit of the communication probabilities,  $\log\left(\frac{P}{1-P}\right)$ , where  $\log$  operates elementwise, is low-rank, reflecting that individuals may have a similar communication pattern across all members of the department. We propose to estimate  $P$  through penalized logistic regression:

$$\min_{\theta \in \mathbb{R}^{N \times N}} \sum_{i,j \in [N] \times [N]} \{\log(\exp(\theta_{ij}) + 1) - X_{ij}\theta_{ij}\} + \lambda \|\theta\|_* \tag{11}$$

In (11),  $\|\cdot\|_*$  denotes the nuclear norm, and its inclusion as a penalty encourages low-rank structure in  $\theta$ . The matrix  $P$  can then be estimated by  $\frac{\exp(\theta)}{1 + \exp(\theta)}$ , where  $\exp$  gives the elementwise exponential. We note that in general  $\text{rank}(\theta) \neq \text{rank}(P)$ , where  $P = \exp(\theta)/(1 + \exp(\theta))$ , so the nuclear norm penalty on  $\theta$  encourages underlying low-rank structure, though this low-rank structure is in  $\theta$ , the logit of  $P$ , instead of  $P$ .

We solve the optimization problem by implementing a proximal Newton algorithm, with a stopping criteria as in Proposition 2 to generate  $\hat{\theta}_{init}$ . We apply Proposition 2 and take  $\hat{\theta}_{init}$  to be a point at which the step length of scaled proximal gradient iterations is less than  $n^{-1/2}$ . We take  $C_n$  to be the Hessian at  $\hat{\theta}_{init}$ , so that  $C_n \rightarrow I_{\theta^*}^{-1}$ .

**Figure 5.** (Color online) (a) The empirical distribution for communication probability between each sender/receiver pair. (b) One step of proximal Newton descent applied to the nuclear-norm penalized logistic regression problem in Section 5.3.



Downloaded from informs.org by [205.155.65.226] on 09 October 2025, at 14:42. For personal use only, all rights reserved.

The logistic regression model that we have formulated satisfies the assumptions of Proposition 1 for  $n$  large enough and penalty parameter  $\lambda$  chosen appropriately, because the regularity conditions (i)–(iii) are satisfied and the minimizer  $\hat{\theta}$  of (11) converges to the true  $\theta^*$  at a  $\sqrt{n}$ -rate (Antoniadis et al. [1]). This permits application of Theorem 3. Theorem 3 gives that the one-step estimator is asymptotically equivalent to the true minimizer of the nuclear-norm penalized logistic regression problem. Figure 5 displays our results, where the plot in (b) depicts the low-rank structure of the solution obtained from applying one step of proximal Newton descent.

### Acknowledgments

The authors thank Johannes Royset for helpful conversations.

### Appendix. Proofs

**Proof of Proposition 2.** Let  $\Delta\hat{\theta}_{init} = \hat{\theta}_{ose} - \hat{\theta}_{init}$ . It is a property of scaled proximal gradient descent that<sup>3</sup>

$$C_n \Delta\hat{\theta}_{init} \in -\nabla g_n(\hat{\theta}_{init}) - \partial h_n(\hat{\theta}_{init} + \Delta\hat{\theta}_{init}).$$

So, for some  $u \in \partial h_n(\hat{\theta}_{init} + \nabla\hat{\theta}_{init})$  and  $v \in \partial h_n(\hat{\theta})$ ,

$$\begin{aligned} & (-\Delta\hat{\theta}_{init})^T C_n (\hat{\theta}_{init} + \Delta\hat{\theta}_{init} - \hat{\theta}) \\ &= (\nabla g_n(\hat{\theta}_{init}) + u)^T (\hat{\theta}_{init} + \Delta\hat{\theta}_{init} - \hat{\theta}) \\ &= (\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta}) + u - v)^T (\hat{\theta}_{init} + \Delta\hat{\theta}_{init} - \hat{\theta}) \end{aligned}$$

By the monotonicity of the subdifferential of a convex function,

$$\geq (\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta}))^T (\hat{\theta}_{init} + \Delta\hat{\theta}_{init} - \hat{\theta}).$$

By the strong convexity of  $f_n$ , with high probability and for some constant  $m$ , we can bound the previous display with

$$\geq \frac{m}{2} \|\hat{\theta}_{init} - \hat{\theta}\|^2 + (\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta}))^T \Delta\hat{\theta}_{init}.$$

We have established

$$(-\Delta\hat{\theta}_{init})^T (C_n (\hat{\theta}_{init} + \Delta\hat{\theta}_{init} - \hat{\theta}) - (\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta}))) \geq \frac{m}{2} \|\hat{\theta}_{init} - \hat{\theta}\|^2.$$

Thus,

$$\begin{aligned} & -(\Delta\hat{\theta}_{init})^T C_n (\Delta\hat{\theta}_{init}) - (\Delta\hat{\theta}_{init})^T C_n (\hat{\theta}_{init} - \hat{\theta}) + (\Delta\hat{\theta}_{init})^T (\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta})) \\ & \geq \frac{m}{2} \|\hat{\theta}_{init} - \hat{\theta}\|^2. \end{aligned}$$

Simplifying,

$$-(\Delta\hat{\theta}_{init})^T C_n (\Delta\hat{\theta}_{init}) + (\Delta\hat{\theta}_{init})^T (\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta}) - C_n (\hat{\theta}_{init} - \hat{\theta})) \geq \frac{m}{2} \|\hat{\theta}_{init} - \hat{\theta}\|^2,$$

so that Cauchy-Schwarz gives

$$\|\Delta\hat{\theta}_{init}\|_{C_n}^2 + \|\Delta\hat{\theta}_{init}\| \|\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta}) - C_n (\hat{\theta}_{init} - \hat{\theta})\| \geq \frac{m}{2} \|\hat{\theta}_{init} - \hat{\theta}\|^2.$$

If  $\lambda_{max}(C_n) < L$ , then

$$L \|\Delta\hat{\theta}_{init}\|^2 + \|\Delta\hat{\theta}_{init}\| \|\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta}) - C_n (\hat{\theta}_{init} - \hat{\theta})\| \geq \frac{m}{2} \|\hat{\theta}_{init} - \hat{\theta}\|^2.$$

We have  $\|\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta})\| \leq M \|\hat{\theta}_{init} - \hat{\theta}\|$  by Lipschitz continuity of  $\nabla g_n$ . Thus,

$$\begin{aligned} & L \|\Delta\hat{\theta}_{init}\|^2 + M \|\Delta\hat{\theta}_{init}\| \cdot \|\hat{\theta}_{init} - \hat{\theta}\| + L \|\Delta\hat{\theta}_{init}\| \cdot \|\hat{\theta}_{init} - \hat{\theta}\| \\ & \geq L \|\Delta\hat{\theta}_{init}\|^2 + \|\Delta\hat{\theta}_{init}\| (M \|\hat{\theta}_{init} - \hat{\theta}\| + \|C_n (\hat{\theta}_{init} - \hat{\theta})\|) \\ & \geq L \|\Delta\hat{\theta}_{init}\|^2 + \|\Delta\hat{\theta}_{init}\| (\|\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta})\| + \|C_n (\hat{\theta}_{init} - \hat{\theta})\|) \\ & \geq L \|\Delta\hat{\theta}_{init}\|^2 + \|\Delta\hat{\theta}_{init}\| \cdot \|\nabla g_n(\hat{\theta}_{init}) - \nabla g_n(\hat{\theta}) - C_n (\hat{\theta}_{init} - \hat{\theta})\| \\ & \geq \frac{m}{2} \|\hat{\theta}_{init} - \hat{\theta}\|^2. \end{aligned}$$

Therefore,

$$L\|\Delta\widehat{\theta}_{init}\|^2 + M\|\Delta\widehat{\theta}_{init}\| \cdot \|\widehat{\theta}_{init} - \widehat{\theta}\| + L\|\Delta\widehat{\theta}_{init}\| \cdot \|\widehat{\theta}_{init} - \widehat{\theta}\| \geq \frac{m}{2}\|\widehat{\theta}_{init} - \widehat{\theta}\|.$$

If  $\|\Delta\widehat{\theta}_{init}\| \geq \|\widehat{\theta}_{init} - \widehat{\theta}\|$ , then

$$\begin{aligned} (2L + M)\|\Delta\widehat{\theta}_{init}\|^2 &\geq \frac{m}{2}\|\widehat{\theta}_{init} - \widehat{\theta}\|^2 \\ \Rightarrow \|\Delta\widehat{\theta}_{init}\| &\geq \sqrt{\frac{m}{2(2L + M)}}\|\widehat{\theta}_{init} - \widehat{\theta}\|. \end{aligned}$$

Otherwise, if  $\|\Delta\widehat{\theta}_{init}\| < \|\widehat{\theta}_{init} - \widehat{\theta}\|$ , then

$$\begin{aligned} (2L + M)\|\Delta\widehat{\theta}_{init}\| \cdot \|\widehat{\theta}_{init} - \widehat{\theta}\| &\geq \frac{m}{2}\|\widehat{\theta}_{init} - \widehat{\theta}\|^2 \\ \Rightarrow \|\Delta\widehat{\theta}_{init}\| &\geq \frac{m}{2(2L + M)}\|\widehat{\theta}_{init} - \widehat{\theta}\|. \end{aligned}$$

Therefore,  $\|\Delta\widehat{\theta}_{init}\| \geq \max\left\{\sqrt{\frac{m}{2(2L+M)}}, \frac{m}{2(2L+M)}\right\}\|\widehat{\theta}_{init} - \widehat{\theta}\|$  is small.

Recalling that  $\Delta\widehat{\theta}_{init} = \widehat{\theta}_{ose} - \widehat{\theta}_{init}$ , the result is proved.  $\square$

**Proof of Proposition 4.** Let  $x^* \in \mathbb{R}^d$ , and let  $w^* \in \text{prox}_{f,C}(x^*)$  be defined as in the statement for  $u^v \rightarrow 0$ , and  $w^v \in \text{prox}_{f,C}(x^* + u^v)$ , with  $w^v \rightarrow w^*$ . We need to show that

$$\lim_{u \rightarrow 0} \frac{e_{cf}(x^* + u) - e_{cf}(x^*) - u^T C(x^* - w^*)}{\|u\|} = 0.$$

By definition of the Moreau envelope,

$$\begin{aligned} e_{cf}(x^* + u^v) &= f(w^v) + \frac{1}{2}(w^v - (x^* + u^v))^T C(w^v - (x^* + u^v)) \\ &\geq f(w^*) + \frac{1}{2}(w^* - (x^* + u^v))^T C(w^* - (x^* + u^v)). \end{aligned}$$

Hence,

$$\begin{aligned} &\frac{e_{cf}(x^* + u^v) - e_{cf}(x^*) - (u^v)^T C(x^* - w^*)}{\|u^v\|} \\ &\leq \frac{f(w^*) + \frac{1}{2}(w^* - (x^* + u^v))^T C(w^* - (x^* + u^v)) - e_{cf}(x^*) - (u^v)^T C(x^* - w^*)}{\|u^v\|} \\ &\leq \frac{\frac{1}{2}(u^v)^T C(w^* - x^*) - (u^v)^T C(x^* - w^*)}{\|u^v\|} \\ &= \frac{(u^v)^T C u^v}{\|u^v\|}, \end{aligned}$$

where the second inequality follows from the definition of  $e_{cf}(x^*) = f(w^*) + \frac{1}{2}(w^* - x^*)^T C(w^* - x^*)$ .

On the other hand, a lower bound on the quotient follows from

$$e_{cf}(x^*) = f(w^*) + \frac{1}{2}(w^* - x^*)^T C(w^* - x^*) \leq f(w^v) + \frac{1}{2}(w^v - x^*)^T C(w^v - x^*).$$

Hence,

$$\begin{aligned} &\frac{e_{cf}(x^* + u^v) - e_{cf}(x^*) - (u^v)^T C(x^* - w^*)}{\|u^v\|} \\ &\geq \frac{e_{cf}(x^* + u^v) - \left(f(w^v) + \frac{1}{2}(w^v - x^*)^T C(w^v - x^*)\right) - (u^v)^T C(x^* - w^*)}{\|u^v\|} \\ &= \frac{\frac{1}{2}(u^v)^T C u^v - (u^v)^T C(w^v - x^*) - (u^v)^T C(x^* - w^*)}{\|u^v\|} \\ &= \frac{\frac{1}{2}(u^v)^T C u^v + (u^v)^T C(w^* - w^v)}{\|u^v\|}. \end{aligned}$$

Combining both bounds, we get the following inequality chain:

$$\begin{aligned} &\frac{\frac{1}{2}(u^v)^T C u^v + (u^v)^T C(w^* - w^v)}{\|u^v\|} \\ &\leq \frac{e_{cf}(x^* + u^v) - e_{cf}(x^*) - (u^v)^T C(x^* - w^*)}{\|u^v\|} \\ &\leq \frac{(u^v)^T C u^v}{\|u^v\|}. \end{aligned}$$

Taking the limit when  $u^v \rightarrow 0$ , the lower bound goes to zero as  $w^v \rightarrow w^*$ , whereas the upper bound also goes to zero as  $v \rightarrow \infty$ . Because this inequality holds for any selection of  $u^v \rightarrow 0$ , this establishes that  $e_{cf}$  is differentiable, with gradient  $\nabla e_{cf}(x^*) = C(x^* - w^*)$ .  $\square$

## Endnotes

<sup>1</sup> This may be shown by using, for example, theorem 10.1 and proposition 8.12 in Rockafellar and Wets [37].

<sup>2</sup> A point  $x^*$  requires inner continuity in the set-convergence sense of variational analysis; see Rockafellar and Wets [37]. If  $\text{prox}_f^c$  is single-valued in a neighborhood of  $x^*$ , then inner-continuity reduces to continuity.

<sup>3</sup> See, for example, Lee et al. [28].

## References

- [1] Antoniadis A, Gijbels I, Nikolova M (2011) Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Ann. Inst. Statist. Math.* 63(3):585–615.
- [2] Bassett R, Deride J (2019) Maximum a posteriori estimators as a limit of bayes estimators. *Math. Programming* 174(1–2):129–144.
- [3] Bauschke HH, Combettes PL (2011) *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Springer, New York).
- [4] Beck A (2017) *First-Order Methods in Optimization* (SIAM, Philadelphia).
- [5] Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* 2(1):183–202.
- [6] Beck A, Teboulle M (2012) Smoothing and first order methods: A unified framework. *SIAM J. Optim.* 22(2):557–580.
- [7] Becker S, Fadili J (2012) A quasi-Newton proximal splitting method. *Adv. Neural Inform. Processing Systems* 25:2618–2626.
- [8] Bickel PJ (1975) One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* 70(350):428–434.
- [9] Boucheron S, Massart P (2011) A high-dimensional wilks phenomenon. *Probab. Theory Related Fields* 150(3–4):405–433.
- [10] Burke JV, Hoheisel T (2013) Epi-convergent smoothing with applications to convex composite functions. *SIAM J. Optim.* 23(3):1457–1479.
- [11] Burke JV, Hoheisel T (2017) Epi-convergence properties of smoothing by infimal convolution. *Set-Valued Variational Anal.* 25(1):1–23.
- [12] Dennis JE, Moré JJ (1974) A characterization of superlinear convergence and its application to quasi-newton methods. *Math. Comput.* 28(126):549–560.
- [13] Eggermont PPB, LaRiccia VN, LaRiccia V (2001) *Maximum Penalized Likelihood Estimation* (Springer, New York).
- [14] Fan J, Chen J (1999) One-step local quasi-likelihood estimation. *J. Royal Statist. Soc. Ser. B. Statist. Methodology* 61(4):927–943.
- [15] Ferguson TS (2017) *A Course in Large Sample Theory* (Routledge, Boca Raton, FL).
- [16] Friedlander MP, Goh G (2017) Efficient evaluation of scaled proximal operators. *Electronic Trans. Numer. Anal.* 46:1–22.
- [17] Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* 33(1):1–22.
- [18] Golub GH, Wilkinson JH (1966) Note on the iterative refinement of least squares solution. *Numer. Math.* 9(2):139–148.
- [19] Hare W, Sagastizábal C (2009) Computing proximal points of nonconvex functions. *Math. Programming* 116(1–2):221–258.
- [20] Hazan E, Levy KY, Shalev-Shwartz S (2016) On graduated optimization for stochastic non-convex problems. *Proc. 33rd Internat. Conf. Machine Learning (JMLR.org)*, 1833–1841.
- [21] Huang C, Huo X (2019) A distributed one-step estimator. *Math. Programming* 174(1–2):41–76.
- [22] Ionides E (2005) Maximum smoothed likelihood estimation. *Statist. Sinica* 15(4):1003–1014.
- [23] Kanzow C, Lechner T (2021) Globalized inexact proximal Newton-type methods for nonconvex composite functions. *Comput. Optim. Appl.* 78(2):377–410.
- [24] Le Cam L (1960) Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.* 3:37–98.
- [25] Le Cam L (1970) On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* 41(3):802–828.
- [26] Le Cam L (1972) *Théorie asymptotique de la décision statistique* (Presses de l’Université de Montréal, Montréal).
- [27] Le Cam L (2012) *Asymptotic Methods in Statistical Decision Theory* (Springer, New York).
- [28] Lee JD, Sun Y, Saunders MA (2014) Proximal Newton-type methods for minimizing composite functions. *SIAM J. Optim.* 24(3):1420–1443.
- [29] Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowledge Discovery Data* 1(1):2-es.
- [30] Luenberger DG, Ye Y (1984) *Linear and Nonlinear Programming* (Springer, Cham, Switzerland).
- [31] Mattingley J, Boyd S (2012) Cvxgen: A code generator for embedded convex optimization. *Optim. Eng.* 13(1):1–27.
- [32] Milzarek A (2016) Numerical methods and second order theory for nonsmooth problems. Unpublished doctoral dissertation, Technische Universität München.
- [33] Mobahi H, Fisher JW (2015) On the link between Gaussian homotopy continuation and convex envelopes. Tai X-C, Bae E, Chan TF, Lysaker M, eds. *Internat. Workshop Energy Minimization Methods Comput. Vision Pattern Recognition* (Springer, Cham, Switzerland), 43–56.
- [34] Parikh N, Boyd S (2014) Proximal algorithms. *Foundations Trends Optim.* 1(3):127–239.
- [35] Pollard D (1997) Another look at differentiability in quadratic mean. Pollard D, Torgersen E, Yang GL, eds. *Festschrift for Lucien Le Cam* (Springer, New York), 305–314.
- [36] Polson NG, Scott JG, Willard BT (2015) Proximal algorithms in statistics and machine learning. *Statist. Sci.* 30(4):559–581.
- [37] Rockafellar RT, Wets RJB (2009) *Variational Analysis* (Springer, Berlin).
- [38] Scheinberg K, Tang X (2016) Practical inexact proximal quasi-newton method with global complexity analysis. *Math. Programming* 160(1–2):495–529.
- [39] Spokoiny V (2012) Parametric estimation. finite sample theory. *Ann. Statist.* 40(6):2877–2909.
- [40] Taddy M (2017) One-step estimator paths for concave regularization. *J. Comput. Graphical Statist.* 26(3):525–536.

- [41] Tseng P, Yun S (2009) A coordinate gradient descent method for nonsmooth separable minimization. *Math. Programming* 117(1):387–423.
- [42] Van der Vaart AW (2000) *Asymptotic Statistics* (Cambridge University Press, Cambridge, UK).
- [43] Xu M, Ye JJ, Zhang L (2015) Smoothing SQP methods for solving degenerate nonsmooth constrained optimization problems with applications to bilevel programs. *SIAM J. Optim.* 25(3):1388–1410.
- [44] Zou H, Li R (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* 36(4):1509–1533.