Approximate Cross-validation: Guarantees for Model Assessment and Selection

Ashia Wilson
Microsoft Research

Maximilian Kasy
Harvard University

Lester Mackey
Microsoft Research

Abstract

Cross-validation (CV) is a popular approach for assessing and selecting predictive models. However, when the number of folds is large, CV suffers from a need to repeatedly refit a learning procedure on a large number of training datasets. Recent work in empirical risk minimization (ERM) approximates the expensive refitting with a single Newton step warm-started from the full training set optimizer. While this can greatly reduce runtime, several open questions remain including whether these approximations lead to faithful model selection and whether they are suitable for non-smooth objectives. We address these questions with three main contributions: (i) we provide uniform non-asymptotic, deterministic model assessment guarantees for approximate CV; (ii) we show that (roughly) the same conditions also guarantee model selection performance comparable to CV; (iii) we provide a proximal Newton extension of the approximate CV framework for non-smooth prediction problems and develop improved assessment guarantees for problems such as ℓ_1-regularized ERM.

1 Introduction

Two important concerns when fitting a predictive model are *model assessment* – estimating the expected performance of the model on a future dataset sampled from the same distribution – and *model selection* – choosing the model hyperparameters to minimize out-of-sample prediction error. Cross-validation (CV) [Geisser, 1975, Stone, 1974] is one of the most widely used techniques for assessment and selection, but it suffers from the need to repeatedly refit a learning procedure on different data subsets.

To reduce the computational burden of CV, recent work proposes to replace the expensive model refitting with an inexpensive surrogate. For example, in the context of regularized empirical risk minimization (ERM), two popular techniques both approximate leave-one-out CV by taking Newton steps from the full-data optimized objective [see, e.g., Beirami et al., 2017, Debruyne et al., 2008, Giordano et al., 2019b, Liu et al., 2014, Rad and Maleki, 2019]. The literature provides single-model guarantees for the assessment quality of these Newton approximations for certain classes of regularized ERM models. Two open questions are whether these approximations are suitable for model selection and whether they are suitable for non-smooth objectives, such as ℓ_1-penalized losses. As put by [Stephenson and Broderick, 2019], "understanding the uses and limitations of approximate CV for selecting λ is one of the most important directions for future work in this area." We address these important open problems in this work.

Our principal contributions are three-fold.

- We provide uniform guarantees for *model assessment* using approximate CV. Specifically, we give conditions which guarantee that the difference between CV and approximate CV is uniformly bounded by a constant of order $1/n^2$, where n is the number of CV folds. In contrast to existing guarantees, our results are non-asymptotic, deterministic, and uniform in λ; our results do not assume a bounded parameter space and provide a more precise convergence rate of $O(1/n^2)$.

- We provide guarantees for *model selection*. We show that roughly the same conditions that guarantee uniform quality assessment results also guarantee that estimators based on parameters tuned by approximate cross-validation and by cross-validation are within $O(1/n)$ distance of each other, so that the approximation error is negligible relative to the sampling variation.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

- We propose a *generalization of approximate CV* that works for general non-smooth penalties. This generalization is based on the proximal Newton method [Lee et al., 2014]. We provide strong model assessment guarantees for this generalization and demonstrate that past non-smooth extensions of ACV fail to satisfy these strong guarantees.

Notation Let $[n] \triangleq \{1, \dots, n\}$, I_d be the $d \times d$ identity matrix, and $\partial\varphi$ denote the subdifferential of a function φ [Rockafellar, 1970]. For any matrix or tensor H, we define $\|H\|_{\mathrm{op}} \triangleq \sup_{v \neq 0 \in \mathbb{R}^d} \|H[v]\|_{\mathrm{op}} / \|v\|_2$ where $\|v\|_{\mathrm{op}} \triangleq \|v\|_2$ is the Euclidean norm. For any Lipschitz vector, matrix, or tensor-valued function f with domain $\mathrm{dom}(f)$, we define the Lipschitz constant $\mathrm{Lip}(f) \triangleq \sup_{x \neq y \in \mathrm{dom}(f)} \frac{\|f(x) - f(y)\|_{\mathrm{op}}}{\|x - y\|_2}$.

2 Cross-validation for Regularized Empirical Risk Minimization

For a given datapoint $z \in \mathcal{X}$ and candidate parameter vector $\beta \in \mathbb{R}^d$, consider the objective function

$$m(z, \beta, \lambda) = \ell(z, \beta) + \lambda \pi(\beta)$$

comprised of a loss function ℓ, a regularizer π, and a regularization parameter $\lambda \in [0, \infty]$. Common examples of loss functions are the least squares loss for regression and the exponential and logistic losses for classification; common examples of regularizers are the ℓ_2^2 (ridge) and ℓ_1 (Lasso) penalties. Our interest is in assessing and selecting amongst estimators fit via *regularized empirical risk minimization (ERM)*:

$$\hat{\beta}(\lambda) \triangleq \begin{cases} \mathrm{argmin}_\beta \, \ell(\mathbb{P}_n, \beta) + \lambda \pi(\beta) & \lambda \in [0, \infty) \\ \mathrm{argmin}_\beta \, \pi(\beta) & \lambda = \infty. \end{cases}$$

Here, $\mathbb{P}_n \triangleq \frac{1}{n} \sum_{i=1}^n \delta_{z_i}$ is an empirical distribution over a given training set with datapoints $z_1, \dots, z_n \in \mathcal{X}$, and we overload notation to write $\ell(\mu, \beta) \triangleq \int \ell(z, \beta) d\mu(z)$ and $m(\mu, \beta, \lambda) \triangleq \ell(\mu, \beta) + \lambda \pi(\beta)$ for any measure μ on $\mathcal{X}$ under which ℓ is integrable.

A standard tool for both model assessment and model selection is the leave-one-out cross-validation (CV)[1] estimate of risk [Geisser, 1975, Stone, 1974]

$$\mathbf{CV}(\lambda) = \tfrac{1}{n} \textstyle\sum_{i=1}^n \ell(z_i, \hat{\beta}_{-i}(\lambda))$$

which is based on the leave-one-out estimators

$$\begin{aligned} \hat{\beta}_{-i}(\lambda) &= \mathrm{argmin}_\beta \, \ell(\mathbb{P}_{n,-i}, \beta) + \lambda \pi(\beta) \\ &= \mathrm{argmin}_\beta \, \tfrac{1}{n} \textstyle\sum_{j \neq i} \ell(z_j, \beta) + \lambda \pi(\beta) \end{aligned} \tag{1}$$

for $\mathbb{P}_{n,-i} \triangleq \frac{1}{n} \sum_{j \neq i} \delta_{z_j}$. Unfortunately, performing leave-one-out CV entails solving an often expensive optimization problem n times for every value of λ evaluated; this makes model selection with leave-one-out CV especially burdensome.

3 Approximating Cross-validation

To provide a faithful estimate of CV while reducing its computational cost, Beirami et al. [2017] (see also [Rad and Maleki, 2019]) considered the following *approximate cross-validation* (ACV) *error*

$$\mathbf{ACV}(\lambda) \triangleq \tfrac{1}{n} \textstyle\sum_{i=1}^n \ell(z_i, \tilde{\beta}_{-i}(\lambda)) \tag{2}$$

based on the approximate leave-one-out CV estimators

$$\tilde{\beta}_{-i}(\lambda) = \hat{\beta}(\lambda) + \nabla_\beta^2 m(\mathbb{P}_{n,-i}, \hat{\beta}(\lambda), \lambda)^{-1} \tfrac{\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))}{n} \tag{3}$$

In effect, **ACV** replaces the task of solving a leave-one-out optimization problem (1) with taking a single Newton step (3) and realizes computational speed-ups when the former is more expensive than the latter. This approximation requires that the objective be everywhere twice-differentiable and therefore does not directly apply to non-smooth ERM problems such as the Lasso. We revisit this issue in Sec. 4.

3.1 Optimizer Comparison

Each ACV estimator (3) can also be viewed as the optimizer of a second-order Taylor approximation to the leave-one-out objective (1), expanded around the full training sample estimate $\hat{\beta}(\lambda)$:

$$\tilde{\beta}_{-i}(\lambda) = \mathrm{argmin}_\beta \; \widehat{m}_2(\mathbb{P}_{n,-i}, \beta, \lambda; \hat{\beta}(\lambda)) \quad \text{for}$$

$$\widehat{m}_2(\mathbb{P}_{n,-i}, \beta, \lambda; \hat{\beta}(\lambda)) \triangleq \sum_{k=0}^{2} \frac{\nabla_\beta^k m(\mathbb{P}_{n,-i}, \hat{\beta}(\lambda), \lambda)[\beta - \hat{\beta}(\lambda)]^{\otimes k}}{k!}.$$

This motivates our optimization perspective on analyzing **ACV**. To understand how well **ACV** approximates **CV** we need only understand how well the optimizers of two related optimization problems approximate one another. As a result, the workhorse of our analysis is the following key lemma, proved in App. A, which controls the difference between the optimizers of similar objective functions. In essence, two optimizers are close if their objectives (or objective gradients) are close and at least one objective has a sharp—that is, not flat—minimum.

[1] We will focus on leave-one-out CV for concreteness, but our results directly apply to any variant of CV, including k-fold and leave-pair-out CV, by treating each fold as a (dependent) datapoint. Leave-pair-out CV is often recommended for AUC estimation [Airola et al., 2009, 2011] but is demanding even for small datasets as $\binom{n}{2}$ folds are required.

Lemma 1 (Optimizer comparison). *Suppose*

$$x_{\varphi_1} \in \operatorname{argmin}_x \varphi_1(x) \quad \text{and} \quad x_{\varphi_2} \in \operatorname{argmin}_x \varphi_2(x).$$

If each φ_i admits an ν_{φ_i} error bound (5), defined in Definition 1 below, then

$$\begin{aligned}&\nu_{\varphi_1}(\|x_{\varphi_1} - x_{\varphi_2}\|_2) + \nu_{\varphi_2}(\|x_{\varphi_1} - x_{\varphi_2}\|_2) \\ &\leq \varphi_2(x_{\varphi_1}) - \varphi_1(x_{\varphi_1}) - (\varphi_2(x_{\varphi_2}) - \varphi_1(x_{\varphi_2})).\end{aligned} \tag{4}$$

If $\varphi_2 - \varphi_1$ is differentiable and φ_2 has ν_{φ_2} gradient growth (6), defined in Definition 1 below, then

$$\nu_{\varphi_2}(\|x_{\varphi_1} - x_{\varphi_2}\|_2) \leq \langle x_{\varphi_1} - x_{\varphi_2}, \nabla(\varphi_2 - \varphi_1)(x_{\varphi_1})\rangle.$$

This result relies on two standard ways of measuring the sharpness of objective function minima:

Definition 1 (Error bound and gradient growth). *Consider the generalized inverse $\nu(r) \triangleq \inf\{s : \omega(s) \geq r\}$ of any non-decreasing function ω with $\omega(0) = 0$. We say a function φ admits an ν* error bound *[Bolte et al., 2017] if*

$$\nu(\|x - x^*\|_2) \leq \varphi(x) - \varphi(x^*) \tag{5}$$

for $x^* = \operatorname{argmin}_{x'} \varphi(x')$ and all $x \in \mathbb{R}^d$. We say a function φ has ν gradient growth *[Nesterov, 2008] if φ is subdifferentiable and*

$$\nu(\|x - y\|_2) \leq \langle y - x, u - v\rangle \tag{6}$$

for all $x, y \in \mathbb{R}^d$ and all $u \in \partial\varphi(y), v \in \partial\varphi(x)$.

Notably, if φ is μ-strongly convex, then φ admits an $\nu_\varphi(r) \triangleq \frac{\mu}{2} r^2$ error bound and $\nu_\varphi(r) \triangleq \mu r^2$ gradient growth, but even non-strongly-convex functions can satisfy quadratic error bounds [Karimi et al., 2016].

3.2 Model Assessment

We now present a deterministic, non-asymptotic approximation error result for **ACV** when used to approximate **CV** for a collection of models indexed by $\lambda \in \Lambda$. Importantly for the model selection results that follow, Thm. 2 shows that the ACV error is an $O(1/n^2)$ approximation to CV error uniformly in λ:

Theorem 2 (**ACV**-**CV** assessment error). *If Assumps. 1 to 3 below hold for some $\Lambda \subseteq [0, \infty]$ and each $(s, r) \in \{(0, 3), (1, 3), (1, 4)\}$, then, for each $\lambda \in \Lambda$,*

$$|\mathbf{ACV}(\lambda) - \mathbf{CV}(\lambda)| \leq \tfrac{\kappa_2}{n^2} \tfrac{\mathrm{B}^\ell_{0,3}}{c_m^2} + \tfrac{\kappa_2}{n^3} \tfrac{\mathrm{B}^\ell_{1,3}}{c_m^3} + \tfrac{\kappa_2^2}{n^4} \tfrac{\mathrm{B}^\ell_{1,4}}{2c_m^4},$$

for $\kappa_p \triangleq \sup_{\lambda \geq 0} \frac{C_{\ell,p+1} + \lambda C_{\pi,p+1}}{p!(c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi])} \leq \max(\frac{C_{p+1,\lambda_\pi}}{p! c_\ell}, \frac{C_{\pi,p+1}}{p! c_\pi})$.

This result, proved in App. B, relies on the following three assumptions:

Assumption 1 (Curvature of objective). *For some $c_\ell, c_\pi, c_m > 0$ and $\lambda_\pi < \infty$, all $i \in [n]$, and all λ, λ' in a given $\Lambda \subseteq [0, \infty]$, $m(\mathbb{P}_{n,-i}, \cdot, \lambda)$ has $\nu_m(r) = c_m r^2$ gradient growth and, for $c_{\lambda',\lambda} \triangleq c_\ell + \lambda' c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]$,*

$$\nabla^2_\beta m(\mathbb{P}_{n,-i}, \hat\beta(\lambda), \lambda') \succeq c_{\lambda',\lambda} \mathrm{I}_d.$$

Assumption 2 (Bounded moments of loss derivatives). *For given $s, r \geq 0$ and $\Lambda \subseteq [0, \infty]$, $\mathrm{B}^\ell_{s,r} < \infty$ where*

$$\mathrm{B}^\ell_{s,r} \triangleq \sup_{\lambda \in \Lambda} \tfrac{1}{n} \textstyle\sum_{i=1}^n \mathrm{Lip}(\nabla_\beta \ell(z_i, \cdot))^s \|\nabla_\beta \ell(z_i, \hat\beta(\lambda))\|_2^r.$$

Assumption 3 (Lipschitz Hessian of objective). *For a given $\Lambda \subseteq [0, \infty]$ and some $C_{\ell,3}, C_{\pi,3} < \infty$,*

$$\mathrm{Lip}(\nabla^2_\beta m(\mathbb{P}_{n,-i}, \cdot, \lambda)) \leq C_{\ell,3} + \lambda C_{\pi,3}, \ \forall \lambda \in \Lambda, i \in [n].$$

Assump. 1 ensures the leave-one-out objectives have curvature near their minima, while Assump. 2 bounds the average discrepancy between the full-data and leave-one-out objectives. Together, Assumps. 1 and 2 ensure that the leave-one-out estimates $\hat\beta_{-i}(\lambda)$ are not too far from the full-data estimate $\hat\beta(\lambda)$ on average. Meanwhile, Assump. 3 ensures that the leave-one-out objective is well-approximated by its second-order Taylor expansion and hence that the ACV estimates $\tilde\beta_{-i}(\lambda)$ are close to the CV estimates $\hat\beta_{-i}(\lambda)$.

Thm. 2 most resembles the (fixed dimension) results obtained by Rad and Maleki [2019, Sec. A.9], who show $|\mathbf{ACV}(\lambda) - \mathbf{CV}(\lambda)| = o_p(1/n)$ under i.i.d. sampling of the datapoints z_i, mild regularity conditions, and the convergence assumptions $\hat\beta(\lambda) \xrightarrow{p} \beta^*(\lambda)$ and $\hat\beta_{-i}(\lambda) \xrightarrow{p} \beta^*(\lambda)$ for some deterministic $\beta^*(\lambda)$. Notably, their guarantees target only the linear prediction setting where $\ell(z_i, \beta) = \phi(y_i, \langle x_i, \beta\rangle)$, and the dependence of the constants in their bound on λ is not discussed.

For each λ, Beirami et al. [2017, Thm. 1] provide an asymptotic, probabilistic analysis of the **ACV** estimators (3) under an assumption that $\hat\beta(\lambda) \xrightarrow{p} \beta^*(\lambda) \in \text{interior}(\Theta)$ where Θ is compact. Specifically, for each value of λ, they guarantee that $\|\hat\beta_{-i}(\lambda) - \tilde\beta_{-i}(\lambda)\|_\infty = O_p(C_\lambda/n^2)$ for a constant C_λ depending on λ in a way that is not discussed. Our Thm. 2 is a consequence of the following similar bound on the estimators employed by cross-validation (1) and approximate cross-validation (3) (see Thm. 14 in App. B): $\|\tilde\beta_{-i}(\lambda) - \hat\beta_{-i}(\lambda)\|_2 \leq \frac{\kappa_2}{c_m^2 n^2} \|\nabla_\beta \ell(z_i, \hat\beta(\lambda))\|_2^2$. In comparison to both [Rad and Maleki, 2019] and [Beirami et al., 2017], our results are non-asymptotic, deterministic, and uniform in λ. They provide a more precise convergence rate than [Rad and Maleki, 2019, Sec. A.9], hold outside of the linear prediction setting, and require no compactness assumptions on the domain of β.

While [Beirami et al., 2017, Rad and Maleki, 2019] assume both a strongly convex objective and a bounded

parameter space, our analysis shows that a separate boundedness assumption on the parameter space is unnecessary; strong convexity alone ensures that $\hat{\beta}(\lambda)$ is uniformly bounded in λ even when the objective and its gradients are unbounded in β. Subsequently, our results apply both to strictly convex objectives (like unregularized logistic regression) when restricted to a compact set and to strongly convex objectives (like ridge-regularized logistic regression) without any domain restrictions.

Moreover, the assumptions underlying Thm. 2 and the other results in this work all hold under standard, easily verified conditions on the objective:

Proposition 3 (Sufficient conditions for assumptions)**.**

1. *Assump. 3 holds for* $\Lambda \subseteq [0,\infty]$ *with* $C_{\pi,3} = \mathrm{Lip}(\nabla^2\pi)$ *and* $C_{\ell,3} = \max_{i\in[n]} \mathrm{Lip}(\nabla^2_\beta \ell(\mathbb{P}_{n,-i},\cdot))$.
2. *If* π *admits an error bound (5) with increasing* ν_π*, and* ℓ *is nonnegative, then*
$$\hat{\beta}(\lambda) \to \hat{\beta}(\infty) \quad as \quad \lambda \to \infty. \tag{7}$$
3. *Assump. 1 holds for* $\Lambda \subseteq [0,\infty]$ *if* $m(\mathbb{P}_{n,-i},\cdot,\lambda)$ *is* c_m*-strongly convex* $\forall \lambda \in \Lambda$ *and* $i \in [n]$*,* π *is strongly convex on a neighborhood of* $\hat{\beta}(\infty)$*, and (7) holds.*
4. *Assump. 2 holds for* $\Lambda \subseteq [0,\infty]$ *and* (s,r) *with* $\mathrm{B}^\ell_{s,r} \le \frac{1}{n}\sum_{i=1}^n L_i^s(\|\nabla_\beta \ell(z_i,\hat{\beta}(\infty))\|_2 + \frac{n-1}{n}\frac{L_i}{c_m}\|\nabla_\beta \ell(\mathbb{P}_n,\hat{\beta}(\infty))\|_2)^r$ *if* $m(\mathbb{P}_{n,-i},\cdot,\lambda)$ *is* c_m*-strongly convex and* $L_i \triangleq \mathrm{Lip}(\nabla_\beta \ell(z_i,\cdot)) < \infty$ *for each* $\lambda \in \Lambda$ *and* $i \in [n]$.

While this result, proved in App. C, is a deterministic statement, it has an immediate probabilistic corollary: if the datapoints $z_1,\dots,z_n$ are i.i.d. draws from a distribution $\mathbb{P}$, then, under the conditions of Prop. 3,

$$C_{\ell,3} \le 3\mathbb{E}_{Z\sim\mathbb{P}}[\mathrm{Lip}(\nabla^2_\beta \ell(Z,\cdot))^4]^{1/4} \quad \text{and}$$
$$\mathrm{B}^\ell_{s,r} \le 2\mathbb{E}_{Z\sim\mathbb{P}}[\mathrm{Lip}(\nabla_\beta \ell(Z,\cdot))^{2s}(\|\nabla_\beta \ell(Z,\hat{\beta}(\infty))\|_2 + \tfrac{1}{c_m}\mathrm{Lip}(\nabla_\beta \ell(Z,\cdot))\|\nabla_\beta \ell(Z,\hat{\beta}(\infty))\|_2)^{2r}]^{1/2}$$

with high probability by Markov's inequality.[2]

3.3 Infinitesimal Jackknife

Giordano et al. [2019b] (see also [Debruyne et al., 2008, Liu et al., 2014]) recently studied a second approximation to leave-one-out cross-validation,

$$\mathbf{ACV}^{\mathrm{IJ}}(\lambda) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(z_i,\tilde{\beta}^{\mathrm{IJ}}_{-i}(\lambda)), \tag{8}$$

based on the *infinitesimal jackknife* (IJ) [Efron, 1982, Jaeckel, 1972] estimate

$$\tilde{\beta}^{\mathrm{IJ}}_{-i}(\lambda) \triangleq \hat{\beta}(\lambda) + \nabla^2_\beta m(\mathbb{P}_n,\hat{\beta}(\lambda),\lambda)^{-1}\frac{\nabla_\beta \ell(z_i,\hat{\beta}(\lambda))}{n}, \tag{9}$$

[2]Note that $\hat{\beta}(\infty) = \mathrm{argmin}_\beta\, \pi(\beta)$ is data-independent.

A potential computational advantage of $\mathbf{ACV}^{\mathrm{IJ}}$ over $\mathbf{ACV}$ is that $\mathbf{ACV}^{\mathrm{IJ}}$ requires only a single Hessian inversion, while $\mathbf{ACV}$ performs n Hessian inversions.[3]

The following theorem, proved in App. D, shows that, under conditions similar to those of Thm. 2, $\mathbf{ACV}$ and $\mathbf{ACV}^{\mathrm{IJ}}$ are nearly the same.

Theorem 4 ($\mathbf{ACV}^{\mathrm{IJ}}$-$\mathbf{ACV}$ assessment error)**.** *If Assumps. 1 and 2 hold for some* $\Lambda \subseteq [0,\infty]$ *and each* $(s,r) \in \{(1,2),(2,2),(3,2)\}$*, then, for each* $\lambda \in \Lambda$*,*

$$|\mathbf{ACV}^{\mathrm{IJ}}(\lambda) - \mathbf{ACV}(\lambda)| \le \frac{\mathrm{B}^\ell_{1,2}}{c^2_{\lambda,\lambda}n^2} + \frac{\mathrm{B}^\ell_{2,2}}{c^3_{\lambda,\lambda}n^3} + \frac{\mathrm{B}^\ell_{3,2}}{2c^4_{\lambda,\lambda}n^4}.$$

Thm. 4 ensures that all of the assessment and selection guarantees for $\mathbf{ACV}$ in this work also extend to $\mathbf{ACV}^{\mathrm{IJ}}$. In particular, Thms. 2 and 4 together imply $\sup_{\lambda\in\Lambda}|\mathbf{ACV}^{\mathrm{IJ}}(\lambda) - \mathbf{CV}(\lambda)| = O(1/n^2)$. A similar $\mathbf{ACV}^{\mathrm{IJ}}$-$\mathbf{CV}$ comparison could be derived from the general infinitesimal jackknife analysis of Giordano et al. [2019b, Cor. 1], which gives $\|\tilde{\beta}^{\mathrm{IJ}}_{-i}(\lambda) - \hat{\beta}_{-i}(\lambda)\|_2 \le C_\lambda/n^2$. However, the constant C_λ in [Giordano et al., 2019b, Cor. 1] is unbounded in λ for non-strongly convex regularizers. Our results only demand curvature from π in the neighborhood of its minimizer and thereby establish $\sup_{\lambda\in\Lambda}|\mathbf{ACV}^{\mathrm{IJ}}(\lambda) - \mathbf{CV}(\lambda)| = O(1/n^2)$ even for robust, non-strongly convex regularizers like the pseudo-Huber penalty [Hartley and Zisserman, 2004, Sec. A6.8], $\pi_\delta(\beta) = \sum_{j=1}^d \delta^2(\sqrt{1+(\beta_j/\delta)^2} - 1)$. In addition, our analyses avoid the compact domain assumption of [Giordano et al., 2019b, Cor. 1].

3.4 Higher-order Approximations to CV

The optimization perspective adopted in this paper naturally points towards generalizations of the estimators (3) and (9). In particular, stronger assessment guarantees can be provided for regularized higher-order Taylor approximations of the objective function. For example, for the regularized p-th order approximation,

$$\mathbf{ACV}_p(\lambda) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(z_i,\tilde{\beta}^{\mathrm{RHO}_p}_{-i}(\lambda)) \quad \text{with}$$
$$\tilde{\beta}^{\mathrm{RHO}_p}_{-i}(\lambda) \triangleq \mathrm{argmin}_\beta\, \widehat{m}_p(\mathbb{P}_{n,-i},\beta,\lambda;\hat{\beta}(\lambda)) + \frac{\mathrm{Lip}(\nabla^p_\beta m(\mathbb{P}_{n,-i},\cdot,\lambda))}{p+1}\|\beta - \hat{\beta}(\lambda)\|_2^{p+1},$$

where $\hat{m}_p$ is a p-th order Taylor expansion of the objective defined by $\hat{f}_p(\beta;\hat{\beta}(\lambda)) \triangleq \sum_{k=0}^p \frac{1}{k!}\nabla^k f(\hat{\beta}(\lambda))[\beta - \hat{\beta}(\lambda)]^{\otimes k}$, we obtain the following improved assessment guarantee, proved in App. B:

Theorem 5 ($\mathbf{ACV}_p$-$\mathbf{CV}$ assessment error)**.** *If Assumps. 1b, 2, and 3b hold for some* $\Lambda \subseteq [0,\infty]$ *and*

[3]Note however that for many losses the Hessians of (3) and (9) differ only by a rank-one update so that all n Hessians can be inverted in time comparable to inverting 1.

each $(s,r) \in \{(0,p+1),(1,p+1),(1,2p)\}$*, then, for* κ_p *defined in Thm. 2 and each* $\lambda \in \Lambda$*,*

$$|\mathbf{ACV}_p(\lambda) - \mathbf{CV}(\lambda)| \leq \frac{2\kappa_p}{n^p}\left(\frac{\mathrm{B}^\ell_{0,p+1}}{c^p_{\lambda,\lambda}} + \frac{\mathrm{B}^\ell_{1,p+1}}{n c^{p+1}_{\lambda,\lambda}} + \frac{\kappa_p \mathrm{B}^\ell_{1,2p}}{n^p c^{2p}_{\lambda,\lambda}}\right).$$

This result relies on the following curvature and smoothness assumptions, which replace Assumps. 1 and 3.

Assumption 1b (Curvature of objective)**.** *For some* $c_\ell, c_\pi > 0$ *and* $\lambda_\pi < \infty$*, all* $i \in [n]$*, and all* λ *in a given* $\Lambda \subseteq [0,\infty]$*,* $m(\mathbb{P}_{n,-i},\cdot,\lambda)$ *has* $\nu_m(r) = c_{\lambda,\lambda} r^2$ *gradient growth for* $c_{\lambda,\lambda} \triangleq c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]$*.*

Assumption 3b (Lipschitz p-th derivative)**.** *For some* $C_{\ell,p+1}, C_{\pi,p+1} < \infty$*, a given* $\Lambda \subseteq [0,\infty]$*, and* $\forall i \in [n]$

$$\mathrm{Lip}(\nabla^p_\beta m(\mathbb{P}_{n,-i},\cdot,\lambda)) \leq C_{\ell,p+1} + \lambda C_{\pi,p+1}, \ \forall \lambda \in \Lambda.$$

Unregularized higher-order IJ approximations to CV were considered in [Debruyne et al., 2008, Liu et al., 2014] and recently analyzed by [Giordano et al., 2019a]. A result similar to Thm. 5 could be derived from [Giordano et al., 2019a, Thm. 1], which controls the approximation error of an *unregularized* IJ version of $\tilde{\beta}^{\mathrm{RHO}_p}_{-i}(\lambda)$, but that work additionally assumes bounded lower-order derivatives. More generally, the framework in App. B provides assessment results for objectives that satisfy weaker curvature conditions than Assump. 1.

3.5 Model Selection

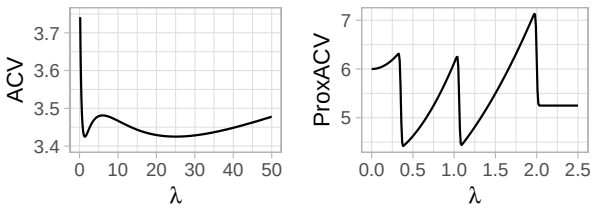
Often, **CV** is used not only to assess a model but also to select a high-quality model for subsequent use. The technique requires training a model with many different values of λ and selecting the one with the lowest **CV** error. If **ACV** is to be used in its stead, we would like to guarantee that the model selected by **ACV** has test error comparable to that selected by **CV**.

When **CV** and **ACV** are uniformly close (as in Thm. 2), we know that any minimizer of **CV** nearly minimizes **ACV** as well, so it suffices to show that all near minimizers of **ACV** have comparable test error. However, this task is made difficult by the potential multimodality of **ACV** and **CV**. As we see in Fig. 1a, even for a benign objective function like the ridge regression objective with quadratic error loss and quadratic penalty, **ACV** and **CV** can have multiple minimizers. Interestingly, our next result, proved in App. F, shows that any near minimizers of **ACV** must produce estimators that are $O(1/\sqrt{n})$ close.

Theorem 6 (**ACV-CV** selection error)**.** *If Assumps. 1 to 3 hold for* $\Lambda \subseteq [0,\infty]$ *and each* $(s,r) \in \{(0,3),(1,2),(1,3),(1,4)\}$*, then* $\forall \lambda', \lambda \in \Lambda$ *with* $\lambda' < \lambda$*,*

$$\|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2^2 \leq \frac{2}{c_m}\left(\frac{4\mathrm{B}^\ell_{0,2}}{c_\ell n} + \Delta\mathbf{ACV} + \frac{\mathrm{B}^\ell_{1,2}}{n^2 c_m^2}\right) \text{ and}$$

$$\|\hat{\beta}(\lambda_{\mathbf{ACV}}) - \hat{\beta}(\lambda_{\mathbf{CV}})\|_2^2 \leq \frac{8}{c_m}\left(\frac{\mathrm{B}^\ell_{0,2}}{c_\ell n} + \frac{A'' c_m^2 + \mathrm{B}^\ell_{1,2}}{4 c_m^2 n^2}\right) \quad (10)$$



(a) Ridge **ACV** (3) (b) Lasso **ProxACV** (13)

Figure 1: **Multimodality of ACV and ProxACV**: (a) **ACV** for $\ell(z,\beta) = (\beta - z)^\top A(\beta - z), \pi(\beta) = \|\beta\|_2^2$, with $A = \mathrm{diag}(1,40)$, sample mean $\bar{z} = \frac{1}{\sqrt{n}}(1.3893, 1.5)$, and sample covariance I_2. (b) **ProxACV** for $\ell(z,\beta) = \frac{1}{2}\|\beta - z\|_2^2, \pi(\beta) = \|\beta\|_1$, with sample mean $\bar{z} = \frac{1}{\sqrt{n}}(\sqrt{1/8}, \sqrt{9/8}, 2)$ and sample covariance I_3.

for $\Delta\mathbf{ACV} \triangleq \mathbf{ACV}(\lambda) - \mathbf{ACV}(\lambda')$*,* $\lambda_{\mathbf{ACV}} \in \mathrm{argmin}_{\lambda\in\Lambda} \mathbf{ACV}(\lambda)$*,* $\lambda_{\mathbf{CV}} \in \mathrm{argmin}_{\lambda\in\Lambda} \mathbf{CV}(\lambda)$*,* $A'' \triangleq 2(\kappa_2 \frac{\mathrm{B}^\ell_{0,3}}{c_m^2} + \frac{\kappa_2}{n}\frac{\mathrm{B}^\ell_{1,3}}{c_m^3} + \frac{\kappa_2^2}{n^2}\frac{\mathrm{B}^\ell_{1,4}}{2c_m^4})$*, and* κ_p *defined in Thm. 2.*

However, the bound (10), established in App. E, only guarantees an approximation error of the same $O(1/\sqrt{n})$ statistical level of the problem and does not fully exploit the $O(1/n^2)$ accuracy provided by the **ACV** estimator (3). Fortunately, we obtain a strengthened $O(1/n)$ guarantee if the objective Hessian is Lipschitz and the minimizers of the loss and regularizer are sufficiently distinct (as measured by $\|\nabla\pi(\hat{\beta}(0))\|_2$).

Theorem 7 (Strong **ACV-CV** selection error)**.** *If Assumps. 1 to 4 hold for some* $\Lambda \subseteq [0,\infty]$ *with* $0 \in \Lambda$ *and each* $(s,r) \in \{(0,3),(1,1),(1,2),(1,3),(1,4)\}$ *and* $\|\nabla\pi(\hat{\beta}(0))\|_2 > 0$*, then for all* $\lambda', \lambda \in \Lambda$ *with* $\lambda' < \lambda$*,*

$$\left|\|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2 - \frac{1}{n}\frac{A}{c_m}\right|^2 \leq \frac{A^2 + 2c_m A'}{n^2 c_m^2} + \frac{2\Delta\mathbf{ACV}}{c_m} \text{ and}$$

$$\left|\|\hat{\beta}(\lambda_{\mathbf{ACV}}) - \hat{\beta}(\lambda_{\mathbf{CV}})\|_2 - \frac{1}{n}\frac{A}{c_m}\right|^2 \leq \frac{A^2 + 2c_m A' + 2c_m A''}{n^2 c_m^2}$$

for $A \triangleq \frac{\mathrm{B}^\ell_{1,1} + \mathrm{B}^\ell_{0,2}\kappa_2}{c_\ell/2} + \frac{\mathrm{B}^\ell_{0,2} C_{\pi,2}\kappa_1^2}{\|\nabla\pi(\hat{\beta}(0))\|_2 c_m}$*,* $A' \triangleq \frac{\mathrm{B}^\ell_{1,2}}{c_m^2}$*,* $\lambda_{\mathbf{ACV}} \in \mathrm{argmin}_{\lambda\in\Lambda} \mathbf{ACV}(\lambda)$*,* $\lambda_{\mathbf{CV}} \in \mathrm{argmin}_{\lambda\in\Lambda} \mathbf{CV}(\lambda)$*, and* $\kappa_p, \Delta\mathbf{ACV}$*, and* A'' *defined in Thms. 2 and 6.*

In fact, Thm. 7, proved in App. F, implies the bound

$$\|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2 = O(\max(1/n, \sqrt{\Delta\mathbf{ACV}})),$$

for *any* values of λ and λ', even if they are not near-minimizers of **ACV**. This result relies on the following additional assumption on the Hessian of the objective, which along with the identifiability condition $\|\nabla\pi(\hat{\beta}(0))\|_2 > 0$ and the curvature of the loss, ensures that two penalty parameters (λ, λ') are close whenever their estimators $(\hat{\beta}(\lambda), \hat{\beta}(\lambda'))$ are close.

Assumption 4 (Bounded Hessian of objective)**.** *For a given* $\Lambda \subseteq [0,\infty]$ *and some* $C_{\ell,2}, C_{\pi,2} < \infty$

$$\nabla^2_\beta m(\mathbb{P}_n, \beta, \lambda) \preccurlyeq (C_{\ell,2} + \lambda C_{\ell,2})\mathrm{I}_d, \quad \forall \lambda \in \Lambda, \beta \in \mathbb{R}^d.$$

Thm. 7 further ensures that the models selected by **CV** and **ACV** have estimators within $O(1/n)$ of one another. Importantly, this approximation error is often negligible compared to the typical $\Omega(1/\sqrt{n})$ statistical estimation error of regularized ERM.

3.6 Failure Modes

One might hope that our **ACV** results extend to objectives that do not meet all of our assumptions, such as the Lasso. For instance, by leveraging the extended definition of an influence function for non-smooth regularized empirical risk minimizers [Avella-Medina et al., 2017], we may define a non-smooth extension of $\mathbf{ACV}^{\mathrm{IJ}}$ that accommodates objectives with undefined Hessians. In the case of squared error loss with an ℓ_1 penalty, $m(\mathbb{P}_n, \beta, \lambda) = \frac{1}{2n}\sum_{i=1}^n \|\beta - z_i\|_2^2 + \lambda\|\beta\|_1$, this amounts to using $\tilde{\beta}_{-i}^{\mathrm{IJ}}(\lambda) \triangleq \hat{\beta}(\lambda) - \frac{1}{n}(\mathbb{I}[\hat{\beta}(\lambda)_j \neq 0](z_{ij} - \hat{\beta}(\lambda)_j))_{j=1}^d$ in the definition (8) of $\mathbf{ACV}^{\mathrm{IJ}}$. Analogous Lasso extensions of **ACV** and $\mathbf{ACV}^{\mathrm{IJ}}$ have been proposed and studied by [Obuchi and Kabashima, 2016, 2018, Rad and Maleki, 2019, Stephenson and Broderick, 2019, Wang et al., 2018]. However, as the following example proved in App. G demonstrates, these extensions do not satisfy the strong uniform assessment and selection guarantees of the prior sections.

Proposition 8. *Suppose* $\ell(z, \beta) = \frac{1}{2}(\beta - z)^2$ *and* $\pi(\beta) = |\beta|$. *Consider a dataset with* $n/4$ *datapoints taking each of the values in* $\{\bar{z} - a, \bar{z} - b, \bar{z} + b, \bar{z} + a\}$ *for* $\bar{z} = \sqrt{2/n}$ *and* $a, b > 0$ *satisfying* $a^2 + b^2 = 2$ *and* $a + b = 2\sqrt{2/\pi}$. *Then* $\lambda = \bar{z}$ *minimizes* $\mathbf{ACV}^{IJ}$ *and* $\mathbf{ACV}^{IJ}(\bar{z}) - \mathbf{CV}(\bar{z}) = \frac{n}{4(n-1)^2}(1 - \frac{4}{\sqrt{n\pi}} + \frac{2}{n})$.

The example in Prop. 8 was constructed to have the same relevant moments as the normal distribution with variance 1 and mean $\sqrt{2/n}$. Notably this $\Omega(1/n)$ assessment error occurs even in the simplest case of $d = 1$; higher-dimensional counterexamples are obtained straightforwardly by creating copies of this example for each dimension. The example demonstrates a failure of deterministic uniform assessment for the Lasso extension of $\mathbf{ACV}^{\mathrm{IJ}}$, and similar counterexamples can be constructed for penalties with well-defined (but non-smooth) second derivatives, like the patched Lasso penalty $\pi(\beta) = \sum_j \min(|\beta_j|, \frac{\delta}{2} + \frac{\beta_j^2}{2\delta})$, for which the standard $\mathbf{ACV}^{\mathrm{IJ}}$ is well-defined.

Proposition 9. *Suppose* $\ell(z, \beta) = \frac{1}{2}(\beta - z)^2$ *and* $\pi(\beta) = \min(|\beta|, \frac{\delta}{2} + \frac{\beta^2}{2\delta})$. *Consider a dataset with* $n/4$ *datapoints taking each of the values in* $\{\bar{z} - a, \bar{z} - b, \bar{z} + b, \bar{z} + a\}$, *where* $\bar{z} = 2\delta$ *and* $a, b > 0$ *satisfy* $a^2 + b^2 = 1$ *and* $a + b = 2\sqrt{2/\pi}$. *Then* $\mathbf{ACV}^{IJ}(\delta) - \mathbf{CV}(\delta) = \delta\sqrt{2/\pi} \cdot \frac{1}{n} + o(\frac{1}{n})$.

The proof of Prop. 9 is contained in App. H. In the following section we propose a modification of **ACV** that addresses these problems.

4 Proximal ACV

Many objective functions involve non-smooth regularizers that violate the assumptions of the preceding section. Common examples are the ℓ_1-regularizer $\pi = \|\cdot\|_1$, often used to engender sparsity for high-dimensional problems, and the elastic net **[add me]**, SLOPE [Bogdan et al., 2013], and nuclear norm [Fazel et al., 2001] penalties. To accommodate non-smooth regularization when approximating CV, several works have proposed either approximating the penalty with a smoothed version [Liu et al., 2018, Rad and Maleki, 2019, Wang et al., 2018] or, for an ℓ_1 penalty, restricting the approximating CV techniques to the support of the full-data estimator [Obuchi and Kabashima, 2016, 2018, Stephenson and Broderick, 2019] as in Sec. 3.6. Experimental evidence with the ℓ_1 penalty suggests these techniques perform well when the support remains consistent across all leave-one-out estimators but can fail otherwise (see [Stephenson and Broderick, 2019, App. D] for an example of failure).

To address the potential inaccuracy of standard **ACV** when coupled with non-smooth regularizers, we recommend use of the proximal operator,

$$\mathrm{prox}_f^H(v) \triangleq \mathrm{argmin}_{\beta \in \mathbb{R}^d} \tfrac{1}{2}\|v - \beta\|_H^2 + f(\beta), \qquad (11)$$

defined for any positive semidefinite matrix H and function f. Specifically, we propose the following *proximal approximate CV error*

$$\mathbf{ProxACV}(\lambda) = \tfrac{1}{n}\sum_{i=1}^n \ell(z_i, \tilde{\beta}_{-i}^{\mathrm{prox}}(\lambda)) \qquad (12)$$

based on the approximate leave-one-out estimators,

$$\begin{aligned}\tilde{\beta}_{-i}^{\mathrm{prox}}(\lambda) &= \mathrm{prox}_{\lambda\pi}^{\mathrm{H}_{\ell,i}}(\hat{\beta}(\lambda) - \mathrm{H}_{\ell,i}^{-1} g_{\ell,i}) \qquad (13)\\ &\triangleq \mathrm{argmin}_{\beta \in \mathbb{R}^d} \tfrac{1}{2}\|\hat{\beta}(\lambda) - \beta\|_{\mathrm{H}_{\ell,i}}^2 + \beta^\top g_{\ell,i} + \lambda\pi(\beta)\end{aligned}$$

with $\mathrm{H}_{\ell,i} = \nabla_\beta^2 \ell(\mathbb{P}_{n,-i}, \hat{\beta}(\lambda))$ and $g_{\ell,i} = \nabla_\beta \ell(\mathbb{P}_{n,-i}, \hat{\beta}(\lambda))$. This estimator optimizes a second-order Taylor expansion of the loss about $\hat{\beta}(\lambda)$ plus the exact regularizer. For many standard objectives, the estimator (13) can be computed significantly more quickly than the exact leave-one-out estimator. Indeed, state-of-the-art solvers like glmnet [Friedman et al., 2010] for ℓ_1-penalized generalized linear models and QUIC [Hsieh et al., 2014] for sparse covariance matrix estimation use a sequence of proximal Newton steps like (13) to optimize their non-smooth objectives. Using **ProxACV** instead entails running these methods for only a single step instead of running them to convergence. In Sec. 5, we give an example of the speed-ups obtainable with this approach.

4.1 Model Assessment

A chief advantage of **ProxACV** is that it is $O(1/n^2)$ close to **CV** uniformly in λ even when the regularizer π lacks the smoothness or curvature previously assumed in Assumps. 1 and 3:

Theorem 10 (**ProxACV-CV** assessment error)**.** *If Assumps. 1c, 2, and 3c hold for $\Lambda \subseteq [0,\infty]$ with $0 \in \Lambda$ and each $(s,r) \in \{(0,3),(1,3),(1,4)\}$, then, $\forall \lambda \in \Lambda$,*

$$|\mathbf{ProxACV}(\lambda) - \mathbf{CV}(\lambda)| \leq \frac{C_{\ell,3}}{n^2}\left(\frac{\mathrm{B}^{\ell}_{0,3}}{2c_m^3} + \frac{\mathrm{B}^{\ell}_{1,3}}{2nc_m^4} + \frac{C_{\ell,3}\mathrm{B}^{\ell}_{1,4}}{8n^2c_m^6}\right).$$

This result, proved in App. I, relies on the following modifications of Assumps. 1 and 3:

Assumption 1c (Curvature of objective)**.** *For $c_m > 0$, all $i \in [n]$, and all λ in a given $\Lambda \subseteq [0,\infty]$, $m(\mathbb{P}_{n,-i},\cdot,\lambda)$ has $\nu_m(r) = c_m r^2$ gradient growth, and π is convex.*

Assumption 3c (Lipschitz Hessian of loss)**.** *For all $i \in [n]$, $\mathrm{Lip}(\nabla^2_\beta \ell(\mathbb{P}_{n,-i},\cdot)) \leq C_{\ell,3} < \infty$.*

Hence, **ProxACV** provides a faithful estimate of **CV** for the non-smooth Lasso, elastic net, SLOPE, and nuclear norm penalties whenever a strongly convex loss with Lipschitz Hessian is used.

4.1.1 Infinitesimal Jackknife

We also propose the following approximation to **CV**,

$$\mathbf{ProxACV}^{\mathrm{IJ}}(\lambda) = \frac{1}{n}\sum_{i=1}^n \ell(z_i, \tilde{\beta}^{\mathrm{prox,IJ}}_{-i}(\lambda)),$$

based on the infinitesimal jackknife-based estimators

$$\tilde{\beta}^{\mathrm{prox,IJ}}_{-i}(\lambda) \triangleq \mathrm{prox}^{\mathrm{H}_\ell}_{\lambda\pi}(\hat{\beta}(\lambda) - \mathrm{H}_\ell^{-1} g_{\ell,i}) \tag{14}$$

with $\mathrm{H}_\ell = \nabla^2_\beta \ell(\mathbb{P}_n, \hat{\beta}(\lambda))$. This approximation is sometimes computationally cheaper than (12) as the same Hessian is used for every estimator. The following result, proved in App. J, shows that **ProxACV** and $\mathbf{ProxACV}^{\mathrm{IJ}}$ are close under our usual assumptions.

Theorem 11 ($\mathbf{ProxACV}^{\mathrm{IJ}}$-**ProxACV** assessment error)**.** *If Assumps. 1c and 2 hold for $\Lambda \subseteq [0,\infty]$ and each $(s,r) \in \{(1,2),(2,2),(3,2)\}$, then for each $\lambda \in \Lambda$,*

$$|\mathbf{ProxACV}(\lambda) - \mathbf{ProxACV}^{\mathrm{IJ}}(\lambda)| \leq \frac{1}{n^2 c_m^2}\mathrm{B}^{\ell}_{1,2} + \frac{1}{2n^4 c_m^4}\mathrm{B}^{\ell}_{3,2} + \frac{1}{n^3 c_m^3}\mathrm{B}^{\ell}_{2,2}.$$

Thms. 10 and 11 imply that $|\mathbf{ProxACV}^{\mathrm{IJ}}(\lambda) - \mathbf{CV}(\lambda)| = O(1/n^2)$ for any $\lambda \in \Lambda$, and subsequently, all assessment and selection guarantees for **ProxACV** in this paper also extend to $\mathbf{ProxACV}^{\mathrm{IJ}}$.

4.2 Model Selection

The following theorem, proved in App. K, establishes a model selection guarantee for **ProxACV**.

Theorem 12 (**ProxACV-CV** selection error)**.** *If Assumps. 1c, 2, and 3c hold for $\Lambda \subseteq [0,\infty]$ and each $(s,r) = \{(0,3),(1,2),(1,3),(1,4)\}$, then $\forall \lambda' < \lambda \in \Lambda$,*

$$\|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2^2 \leq \frac{2}{nc_m}\left(\frac{4\mathrm{B}^{\ell}_{0,2}}{c_m} + \frac{\mathrm{B}^{\ell}_{1,2}}{nc_m^2} + \mathbf{\Delta ProxACV}\right)$$

and $\|\hat{\beta}(\lambda_{\mathbf{PACV}}) - \hat{\beta}(\lambda_{\mathbf{CV}})\|_2^2 \leq \frac{2}{nc_m}\left(\frac{4\mathrm{B}^{\ell}_{0,2}}{c_m} + \frac{\mathrm{B}^{\ell}_{1,2}}{nc_m^2} + \tilde{A}\right)$

for $\lambda_{\mathbf{CV}} \in \mathrm{argmin}_{\lambda\in\Lambda}\mathbf{CV}(\lambda)$, $\lambda_{\mathbf{PACV}} \in \mathrm{argmin}_{\lambda\in\Lambda}\mathbf{ProxACV}(\lambda)$ and $\tilde{A} \triangleq \frac{C_{\ell,3}}{n^2}\left(\frac{\mathrm{B}^{\ell}_{0,3}}{c_m^3} + \frac{\mathrm{B}^{\ell}_{1,3}}{nc_m^4} + \frac{C_{\ell,3}\mathrm{B}^{\ell}_{1,4}}{4n^2c_m^6}\right)$.

Notably and unlike the **ACV** selection results of Thms. 6 and 7, Thm. 12 demands no curvature or smoothness from the regularizer. Moreover, for ℓ_1-penalized problems, the $O(1/\sqrt{n})$ error bound is tight as the following example illustrates.

Proposition 13. *Suppose $\ell(z,\beta) = \frac{1}{2}(\beta - z)^2$ and $\pi(\beta) = |\beta|$. Consider a dataset evenly split between the values $a = \sqrt{2}$ and $b = 2\sqrt{2/n} - \sqrt{2}$ for $n \geq 4$. Then $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i = \sqrt{2/n}$, and $\mathbf{ProxACV}(0) - \mathbf{ProxACV}(\bar{z}) = \frac{5}{2n^2}$, but $\hat{\beta}(0) - \hat{\beta}(\bar{z}) = \bar{z} - 0 = \sqrt{2/n}$.*

The proof of this proposition is contained in App. L. At the heart of this counterexample is multimodality, which can occur for ℓ_1 penalized objectives (see Fig. 1b), much as it did for the ridge example of Fig. 1a. In particular, for ℓ_1 regularized objectives, the modes of **ACV** and **ProxACV** can be $\Omega(1/\sqrt{n})$ apart. While this example prevents us from obtaining an $O(1/n)$ deterministic model selection bound for **ProxACV** in the worst case, it is possible that Thm. 12 can be generically strengthened (as in Thm. 7) when the the minimizers of the loss and regularizer are sufficiently separated. In addition, the possibility of a strong *probabilistic* model selection bound is not precluded.

5 Experiments

We present two sets of experiments to illustrate the value of the newly proposed **ProxACV** procedure. The first compares the assessment quality of **ProxACV** and prior non-smooth **ACV** proposals. The second compares the speed of **ProxACV** to exact **CV**. See `https://github.com/aswilson07/ApproximateCV` for code reproducing all experiments.

5.1 ProxACV versus ACV and $\mathbf{ACV}^{\mathrm{IJ}}$

To compare **ProxACV** with prior non-smooth extensions of **ACV** and $\mathbf{ACV}^{\mathrm{IJ}}$, we adopt the code and the ℓ_1-regularized logistic regression experimental setup of Stephenson and Broderick [2019, App. F]. We use the $\beta \in \mathbb{R}^{151}$ setting, changing only the number of datapoints to $n = 150$ (see App. M for more details). For two ranges of λ values, we compare exact **CV** with

the approximations provided by **ProxACV** (13) and the prior non-smooth extensions of **ACV** and **ACV$^{\text{IJ}}$** discussed in Sec. 3.6 and detailed in App. M. Fig. 2 (top) shows that for sufficiently large λ all three approximations closely match **CV**. However, as noted in [Stephenson and Broderick, 2019, App. F], the non-smooth extension of **ACV$^{\text{IJ}}$** provides an extremely poor approximation leading to grossly incorrect model selection as λ decreases. Moreover, the approximation provided by the non-smooth extension of **ACV** also deteriorates as λ decreases; this is especially evident in the small λ range of Fig. 2 (bottom), where the relative error of the **ACV** approximation exceeds 100%. Meanwhile, **ProxACV** provides a significantly more faithful approximation of **CV** across the range of large and small λ values.

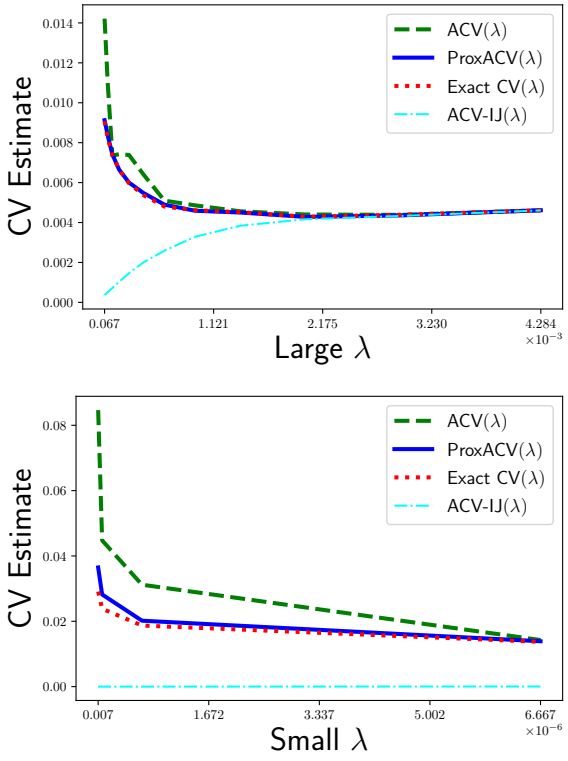


Figure 2: **ProxACV vs. ACV and ACV$^{\text{IJ}}$:** Fidelity of non-smooth **CV** approximations in the ℓ_1-regularized logistic regression setup of Sec. 5.1.

5.2 ProxACV Speed-up

We next benchmark the speed-up of **ProxACV** over **CV** on the task of sparse inverse covariance estimation. using three biological data sets preprocessed by Li and Toh [2010]: Arabidopsis ($p = 834$, $n = 118$), Leukemia ($p = 1,225$, $n = 72$), and Lymph ($p = 587$, $n = 148$). We employ the standard graphical Lasso objective for matrices $\beta \in \mathbb{R}^{p \times p}$ (see App. M for details) and compute our **CV** and full-data estimators using the released Matlab implementation of the state-of-the-art graphical Lasso solver, QUIC [Hsieh et al., 2014]. Since QUIC optimizes $m(\mathbb{P}_{n,-i}, \beta, \lambda)$ using a proximal Newton algorithm, we compute our proximal ACV estimators by running QUIC for a single proximal Newton step instead of running it to convergence. We follow the exact experimental setup of [Hsieh et al., 2014, Fig. 2] which employs a penalty of $\lambda = 0.5$ for all datasets. The timing for each leave-one-out iteration of **CV** and **ProxACV** was computed using a single core on a 2.10 GHz Intel Xeon E5-4650 CPU. In Fig. 3, we display the average relative error, $1 - \textbf{ProxACV}(\lambda)/\textbf{CV}(\lambda)$, and running time ($\pm$ 1 standard deviation) over 10 independent runs. We see that **ProxACV** delivers 14 - 27-fold average speed-ups over **CV** with relative errors below 0.02 in each case.

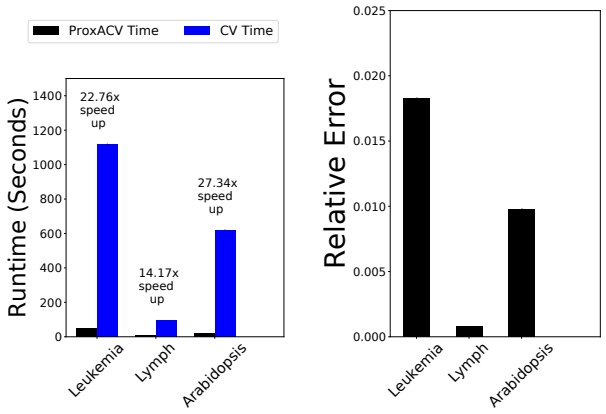


Figure 3: **ProxACV vs. CV:** Speed-up and relative error of **ProxACV** over exact **CV** on three biological datasets using QUIC sparse inverse covariance estimation (see Sec. 5.2). **ProxACV** provides a faithful estimate of **CV** with significant computational gains.

Importance of curvature Thm. 10 relies on the curvature c_m of the objective, and, in general, such a curvature assumption is necessary for **ProxACV** to provide a faithful approximation. The graphical Lasso objective is strictly but not strongly convex, but the default λ choice of [Hsieh et al., 2014] effectively limits the domain of m to a compact set with a sizable curvature. However, as λ decreases, the effective domain of m grows, and the curvature decays leading to a worse approximation. For example, when $\lambda = 0.25$ on the Arabidopsis dataset, we obtain a 97.43-fold average speed-up but with 0.137 mean relative error.

Acknowledgments

We thank Kim-Chuan Toh, Matyas Sustik, and Cho-Jui Hsieh for sharing their covariance estimation data

and Will Stephenson for sharing his approximate cross-validation code. We also thank the anonymous reviewers for their role in improving this manuscript. Special thanks to Gary Chamberlain who inspired this project – Rest in peace.

References

A. Airola, T. Pahikkala, W. Waegeman, B. D. Baets, and T. Salakoski. A comparison of auc estimators in small-sample studies. In S. Dzeroski, P. Guerts, and J. Rousu, editors, *Proceedings of the third International Workshop on Machine Learning in Systems Biology*, volume 8 of *Proceedings of Machine Learning Research*, pages 3–13, Ljubljana, Slovenia, 05–06 Sep 2009. PMLR. URL `http://proceedings.mlr.press/v8/airola10a.html`.

A. Airola, T. Pahikkala, W. Waegeman, B. D. Baets, and T. Salakoski. An experimental comparison of cross-validation techniques for estimating the area under the roc curve. *Computational Statistics & Data Analysis*, 55(4):1828 – 1844, 2011. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2010.11.018. URL `http://www.sciencedirect.com/science/article/pii/S0167947310004469`.

M. Avella-Medina et al. Influence functions for penalized m-estimators. *Bernoulli*, 23(4B):3178–3196, 2017.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh. On optimal generalizability in parametric learning. In *Advances in Neural Information Processing Systems*, pages 3455–3465, 2017.

M. Bogdan, E. v. d. Berg, W. Su, and E. Candes. Statistical estimation and testing via the sorted l1 norm. *arXiv preprint arXiv:1310.1969*, 2013.

J. Bolte, T.-P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471507, 2017.

M. Debruyne, M. Hubert, and J. A. Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9 (Oct):2377–2400, 2008.

B. Efron. *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam, 1982.

M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 6, pages 4734–4739. IEEE, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

R. Giordano, M. I. Jordan, and T. Broderick. A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019a.

R. Giordano, W. Stephenson, R. Liu, M. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1139–1147, 16–18 Apr 2019b.

R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

C. Hsieh, M. Sustik, I. Dhillon, and P. Ravikumar. Quic: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15:2911–2947, 10 2014. ISSN 1532-4435.

L. Jaeckel. The infinitesimal jackknife, memorandum. Technical report, MM 72-1215-11, Bell Lab. Murray Hill, NJ, 1972.

H. Karimi, J. Nutini, and M. W. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, pages 795–811, 2016.

J. D. Lee, Y. Sun, and M. A. Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

L. Li and K.-C. Toh. An inexact interior point method for l1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2(3):291–315, Dec 2010.

Y. Liu, S. Jiang, and S. Liao. Efficient approximation of cross-validation for kernel methods using bouligand influence function. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 324–332, Bejing, China, 22–24 Jun 2014. PMLR. URL `http://proceedings.mlr.press/v32/liua14.html`.

Y. Liu, H. Lin, L. Ding, W. Wang, and S. Liao. Fast cross-validation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial*

Intelligence, IJCAI-18, pages 2497–2503. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

Y. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008. ISSN 0025-5610.

Y. Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, pages 1–27, 2019.

T. Obuchi and Y. Kabashima. Cross validation in lasso and its acceleration. *Journal of Statistical Mechanics*, 2016.

T. Obuchi and Y. Kabashima. Accelerating cross-validation in multinomial logistic regression with l1-regularization. *Journal of Machine Learning Research*, 2018.

K. R. Rad and A. Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out. *arXiv preprint arXiv:1801.10243*, 2019.

R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

W. T. Stephenson and T. Broderick. Sparse approximate cross-validation for high-dimensional glms. *arXiv preprint arXiv:1905.13657*, 2019.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

S. Wang, W. Zhou, H. Lu, A. Maleki, and V. Mirrokni. Approximate leave-one-out for fast parameter tuning in high dimensions. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5228–5237, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.