A Proof of Lemma 1: Optimizer comparison

The first claim (4) follows immediately from the definition of the error bound (5).

To establish the second claim, we note that our (sub)differentiability assumptions and the optimality of x_{φ_1} and x_{φ_2} imply that $0 \in \partial\varphi_2(x_{\varphi_2})$ and $0 = u + \nabla(\varphi_1 - \varphi_2)(x_{\varphi_1})$ for some $u \in \partial\varphi_2(x_{\varphi_1})$. Gradient growth (6) now implies

$$\nu_{\varphi_2}(\|x_{\varphi_1} - x_{\varphi_2}\|_2) \leq \langle x_{\varphi_1} - x_{\varphi_2}, u - 0\rangle = \langle x_{\varphi_1} - x_{\varphi_2}, \nabla(\varphi_2 - \varphi_1)(x_{\varphi_1})\rangle.$$

B Proof of Thms. 2 and 5: ACV-CV and ACV$_p$-CV assessment error

Thms. 2 and 5 will follow from the following more detailed statement, proved in App. B.1. Consider the higher-order gradient estimator

$$\mathbf{ACV}_p^{\mathrm{HO}}(\lambda) \triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(z_i, \tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda)) \quad \text{with} \quad \tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) \triangleq \operatorname{argmin}_\beta \widehat{m}_p(\mathbb{P}_{n,\text{-}i}, \beta, \lambda; \hat{\beta}(\lambda)),$$

which recovers our approximate CV error (2) and estimate (3) when $p = 2$. We will make use of the following assumptions which generalize Assumps. 1 and 1b.

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

Assumption 1e (Curvature of Taylor approximation). *For some $p, q, c_\ell, c_\pi > 0$ and $\lambda_\pi < \infty$, all $i \in [n]$, and all λ in a given $\Lambda \subseteq [0, \infty]$, $\widehat{m}_p(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda))$ has $\nu(r) = c_{\lambda,\lambda} r^q$ gradient growth, where $c_{\lambda,\lambda} \triangleq c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]$.*

Assumption 1f (Curvature of regularized Taylor approximation). *For some $p, q, c_\ell, c_\pi > 0$ and $\lambda_\pi < \infty$, all $i \in [n]$, and all λ in a given $\Lambda \subseteq [0, \infty]$, $\widehat{m}_p(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda)) + \frac{\mathrm{Lip}(\nabla_\beta^p m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda))}{p+1}\|\cdot - \hat{\beta}(\lambda)\|_2^{p+1}$ has $\nu(r) = c_{\lambda,\lambda} r^q$ gradient growth, where $c_{\lambda,\lambda} \triangleq c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]$.*

Theorem 14 (**ACV$_p$-CV** and **ACV$_p^{\mathrm{HO}}$-CV** assessment error). *If Assump. 1d holds for some $\Lambda \subseteq [0, \infty]$, then, for all $\lambda \in \Lambda$ and $i \in [n]$,*

$$\|\hat{\beta}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2^{q-1} \leq \tfrac{1}{n}\tfrac{1}{c_m}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2. \tag{15}$$

If Assumps. 3b, 1d, and 1e hold for some $\Lambda \subseteq [0, \infty]$, then, for all $\lambda \in \Lambda$ and $i \in [n]$,

$$\|\tilde{\beta}_{\text{-}i}^{HO_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2^{q-1} \leq \kappa_{p,\lambda}^\lambda \|\hat{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2^p \tag{16a}$$

for $\kappa_{p,\lambda}^\lambda \triangleq \frac{C_{\ell,p+1} + \lambda C_{\pi,p+1}}{p!(c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi])}$.

If Assumps. 3b, 1d, and 1f hold for some $\Lambda \subseteq [0, \infty]$, then, for all $\lambda \in \Lambda$ and $i \in [n]$,

$$\|\tilde{\beta}_{\text{-}i}^{RHO_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2^{q-1} \leq 2\kappa_{p,\lambda}^\lambda \|\hat{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2^p. \tag{16b}$$

If Assumps. 2, 3b, 1d, and 1e hold for some $\Lambda \subseteq [0, \infty]$ and each $(s, r) \in \{(0, \frac{p+(q-1)^2}{(q-1)^2}), (1, \frac{2p}{(q-1)^2}), (1, \frac{p+q-1}{(q-1)^2})\}$, then, for all $\lambda \in \Lambda$,

$$\begin{aligned} &|\mathbf{ACV}_p^{HO}(\lambda) - \mathbf{CV}(\lambda)| \\ &\leq \frac{1}{n^{\frac{p}{(q-1)^2}}} \frac{(\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}}}{c_m^{\frac{p}{(q-1)^2}}} \mathrm{B}^\ell_{0,\frac{p+(q-1)^2}{(q-1)^2}} + \frac{1}{2}\frac{1}{n^{\frac{2p}{(q-1)^2}}} \frac{(\kappa_{p,\lambda}^\lambda)^{\frac{2}{q-1}}}{c_m^{\frac{2p}{(q-1)^2}}} \mathrm{B}^\ell_{1,\frac{2p}{(q-1)^2}} + \frac{1}{n^{\frac{p+q-1}{(q-1)^2}}} \frac{(\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}}}{c_m^{\frac{p+q-1}{(q-1)^2}}} \mathrm{B}^\ell_{1,\frac{p+q-1}{(q-1)^2}} \quad \textit{and} \end{aligned} \tag{17a}$$

If Assumps. 2, 3b, 1d, and 1f hold for some $\Lambda \subseteq [0, \infty]$ and each $(s, r) \in \{(0, \frac{p+(q-1)^2}{(q-1)^2}), (1, \frac{2p}{(q-1)^2}), (1, \frac{p+q-1}{(q-1)^2})\}$, then, for all $\lambda \in \Lambda$,

$$\begin{aligned} &|\mathbf{ACV}_p(\lambda) - \mathbf{CV}(\lambda)| \\ &\leq \frac{1}{n^{\frac{p}{(q-1)^2}}} \frac{(2\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}}}{c_m^{\frac{p}{(q-1)^2}}} \mathrm{B}^\ell_{0,\frac{p+(q-1)^2}{(q-1)^2}} + \frac{1}{2}\frac{1}{n^{\frac{2p}{(q-1)^2}}} \frac{(2\kappa_{p,\lambda}^\lambda)^{\frac{2}{q-1}}}{c_m^{\frac{2p}{(q-1)^2}}} \mathrm{B}^\ell_{1,\frac{2p}{(q-1)^2}} + \frac{1}{n^{\frac{p+q-1}{(q-1)^2}}} \frac{(2\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}}}{c_m^{\frac{p+q-1}{(q-1)^2}}} \mathrm{B}^\ell_{1,\frac{p+q-1}{(q-1)^2}}. \end{aligned} \tag{17b}$$

Thm. 2 follows from Thm. 14 with $p = q = 2$ since Assump. 1 implies $\mu = c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]$ strong convexity and hence $\nu(r) = \mu r^2$ gradient growth for each $\widehat{m}_2(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda))$.

Thm. 5 follows from Thm. 14 with $q = 2$ since Assumps. 1b and 3b and the following lemma imply that each $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$ and $\widehat{m}_p(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda)) + \frac{\text{Lip}(\nabla_\beta^p m(\mathbb{P}_{n,\text{-}i},\cdot,\lambda))}{p+1} \| \cdot - \hat{\beta}(\lambda)\|_2^{p+1}$ has $\mu = c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]$ strong convexity and hence $\nu(r) = \mu r^2$ gradient growth.

Lemma 15 (Curvature of regularized Taylor approximation)**.** *If φ is μ strongly convex and $\nabla^p \varphi$ is Lipschitz, then $\Phi(x) \triangleq \widehat{\varphi}_p(x; w) + \frac{\text{Lip}(\nabla^p \varphi)}{(p+1)!} \|x - w\|_2^{p+1}$ is μ strongly convex.*

Proof This result is inspired by [Nesterov, 2019, Thm. 1]. In particular, by Taylor's theorem with integral remainder, we can bound the residual between a function and its Taylor approximation as

$$|\varphi(x) - \widehat{\varphi}_p(x; w)| \leq \tfrac{\text{Lip}(\nabla^p \varphi)}{(p+1)!} \|x - w\|_2^{p+1}$$

Note also that for $d(x) = \frac{1}{p}\|x\|^p$

$$\nabla^2 d(x) = (p-2)\|x\|^{p-4} x x^\top + \|x\|^{p-2} \mathrm{I}_d \succcurlyeq \|x\|^{p-2} \mathrm{I}_d. \tag{18}$$

For $p \geq 2$, applying the same reasoning to $\langle \nabla f(\cdot), h\rangle$ and $\langle \nabla^2 f(\cdot) h, h\rangle$ we can similarly conclude:

$$\|\nabla \varphi(x) - \nabla \widehat{\varphi}_p(x; w)\|_{\text{op}} \leq \tfrac{\text{Lip}(\nabla^p \varphi)}{p!} \|x - w\|_2^p$$
$$\|\nabla^2 \varphi(x) - \nabla^2 \widehat{\varphi}_p(x; w)\|_{\text{op}} \leq \tfrac{\text{Lip}(\nabla^p \varphi)}{(p-1)!} \|x - w\|_2^{p-1}$$

Subsequently, for any direction $h \in \mathbb{R}^d$

$$\langle (\nabla^2 \varphi(x) - \nabla^2 \widehat{\varphi}_p(x; w)) h, h\rangle \leq \|\nabla^2 \varphi(x) - \nabla^2 \widehat{\varphi}_p(x; w)\|_{\text{op}} \cdot \|h\|_2^2 \leq \tfrac{\text{Lip}(\nabla^p \varphi)}{(p-1)!} \|x - w\|_2^{p-1} \cdot \|h\|_2^2,$$

and therefore,

$$\nabla^2 \varphi(x) \preccurlyeq \nabla^2 \widehat{\varphi}_p(x; w) + \tfrac{\text{Lip}(\nabla^p \varphi)}{(p-1)!} \|x - w\|_2^{p-1} \mathrm{I}_d \overset{(18)}{\preccurlyeq} \nabla^2 \Phi(x).$$

□

B.1 Proof of Thm. 14: ACV$_p$-CV and ACV$_p^{\text{HO}}$-CV assessment error

B.1.1 Proof of (15): Proximity of CV and full-data estimators

We begin with a lemma that translates the polynomial gradient growth of our objective into a bound on the difference between a full-data estimator $\hat{\beta}(\lambda)$ and a leave-one-out estimator $\hat{\beta}_{\text{-}i}(\lambda)$.

Lemma 16 (Proximity of CV and full-data estimators)**.** *Fix any $\lambda \in [0, \infty)$ and $i \in [n]$. If $\ell(z_i, \cdot)$ is differentiable, and $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$ has $\nu_m(r) = c_m r^q$ gradient growth (6) for $c_m > 0$ and $q > 0$, then*

$$\|\hat{\beta}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2^{q-1} \leq \tfrac{1}{n} \tfrac{1}{c_m} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2.$$

Proof The result follows from the Optimizer Comparison Lemma 1 with $\varphi_1(\beta) = m(\mathbb{P}_n, \beta, \lambda)$ and $\varphi_2(\beta) = m(\mathbb{P}_{n,\text{-}i}, \beta, \lambda)$ and Cauchy-Schwarz, as

$$\begin{aligned} c_m \|\hat{\beta}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2^q &\leq \langle \hat{\beta}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda), \nabla_\beta m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda) - \nabla_\beta m(\mathbb{P}_n, \hat{\beta}(\lambda), \lambda)\rangle \\ &= \tfrac{1}{n} \langle \hat{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda), \nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\rangle \leq \tfrac{1}{n} \|\hat{\beta}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2 \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2. \end{aligned}$$

□

Now fix any $\lambda \in \Lambda$ and $i \in [n]$. If $\lambda = \infty$, then $\hat{\beta}(\lambda) = \hat{\beta}_{\text{-}i}(\lambda)$, ensuring the result (15). If $\lambda \neq \infty$, then our assumptions and Lemma 16 immediately establish the result (15).

B.1.2 Proof of (16a): Proximity of $\mathrm{ACV}_p^{\mathrm{HO}}$ and CV estimators

The result (16a) will follow from a general Taylor comparison lemma that bounds the optimizer error introduced by approximating part of an objective with its Taylor polynomial.

Lemma 17 (Taylor comparison)**.** *Suppose*

$$x_\varphi \in \operatorname*{argmin}_x \varphi(x) + \varphi_0(x) \quad \textit{and} \quad x_{\widehat{\varphi}_p} \in \operatorname*{argmin}_x \widehat{\varphi}_p(x; w) + \varphi_0(x).$$

for $\widehat{\varphi}_p(x; w) \triangleq \sum_{i=0}^{p} \frac{1}{i!} \nabla^i \varphi(w)[x - w]^{\otimes i}$ *the p-th-order Taylor polynomial of* φ *about a point* w*. If* $\nabla^p \varphi$ *is Lipschitz and* $\widehat{\varphi}_p(\cdot; w) + \varphi_0$ *has* $\nu(r) = \mu r^q$ *gradient growth* (6) *for* $\mu > 0$ *and* $q > 0$*, then*

$$\|x_\varphi - x_{\widehat{\varphi}_p}\|_2^{q-1} \leq \tfrac{\mathrm{Lip}(\nabla^p \varphi)}{\mu} \tfrac{1}{p!} \|x_\varphi - w\|_2^p.$$

Proof Define $f(x) = \langle x_{\widehat{\varphi}_p} - x_\varphi, \nabla \varphi(x) \rangle$. The result follows from the Optimizer Comparison Lemma 1 with $\varphi_1 = \varphi + \varphi_0$ and $\varphi_2 = \widehat{\varphi}_p(\cdot; w) + \varphi_0$, Taylor's theorem with integral remainder, and Cauchy-Schwarz as

$$\begin{aligned}\mu\|x_\varphi - x_{\widehat{\varphi}_p}\|_2^q &\leq \langle x_\varphi - x_{\widehat{\varphi}_p}, \nabla_x \widehat{\varphi}_p(x_\varphi; w) - \nabla \varphi(x_\varphi) \rangle = f(x_\varphi) - \textstyle\sum_{i=0}^{p-1} \frac{1}{i!} \nabla^i f(w)[x_\varphi - w]^{\otimes i} \\ &\leq \tfrac{\mathrm{Lip}(\nabla^{p-1} f)}{p!} \|x_\varphi - w\|_2^p \leq \|x_\varphi - x_{\widehat{\varphi}_p}\|_2 \tfrac{\mathrm{Lip}(\nabla^p \varphi)}{p!} \|x_\varphi - w\|_2^p.\end{aligned}$$

□

To see this, fix any $\lambda \in \Lambda$ and $i \in [n]$, and consider the choices $\varphi = m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$, $\varphi_0 \equiv 0$, and $w = \hat{\beta}(\lambda)$. By Assump. 1e, $\widehat{\varphi}_p(\cdot; w) + \varphi_0$ has $\nu(r) = \mu r^q$ gradient growth for $\mu = c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]$. Since $\mathrm{Lip}(\nabla^p \varphi) \leq C_{\ell,p+1} + \lambda C_{\pi,p+1}$ by Assump. 3b, the desired result (16a) follows from Lemma 17.

B.1.3 Proof of (16b): Proximity of ACV_p and CV estimators

The result (16b) will follow from a regularized Taylor comparison lemma that bounds the optimizer error introduced by approximating part of an objective with a regularized Taylor polynomial.

Lemma 18 (Regularized Taylor comparison)**.** *Suppose*

$$x_\varphi \in \operatorname{argmin}_x \varphi(x) + \varphi_0(x) \quad \textit{and} \quad x_{\widehat{\varphi}_p} \in \operatorname{argmin}_x \widehat{\varphi}_p(x; w) + \tfrac{\mathrm{Lip}(\nabla^p \varphi)}{(p+1)!} \|x - w\|_2^{p+1} + \varphi_0(x).$$

for $\widehat{\varphi}_p(x; w) \triangleq \sum_{i=0}^{p} \frac{1}{i!} \nabla^i \varphi(w)[x - w]^{\otimes i}$ *the p-th-order Taylor polynomial of* φ *about a point* w*. If* $\nabla^p \varphi$ *is Lipschitz and* $\widehat{\varphi}_p(\cdot; w) + \frac{\mathrm{Lip}(\nabla^p \varphi)}{(p+1)!} \| \cdot - w\|_2^{p+1} + \varphi_0$ *has* $\nu(r) = \mu r^q$ *gradient growth* (6) *for* $\mu > 0$ *and* $q > 0$*, then*

$$\|x_\varphi - x_{\widehat{\varphi}_p}\|_2^{q-1} \leq \tfrac{2\,\mathrm{Lip}(\nabla^p \varphi)}{\mu} \tfrac{1}{p!} \|x_\varphi - w\|_2^p.$$

Proof Define $f(x) = \langle x_{\widehat{\varphi}_p} - x_\varphi, \nabla \varphi(x) \rangle$. The result follows from the Optimizer Comparison Lemma 1 with $\varphi_1 = \varphi + \varphi_0$ and $\varphi_2 = \widehat{\varphi}_p(\cdot; w) + \frac{\mathrm{Lip}(\nabla^p \varphi)}{(p+1)!} \| \cdot - w\|_2^{p+1} + \varphi_0$, Taylor's theorem with integral remainder, and Cauchy-Schwarz as

$$\begin{aligned}\mu\|x_\varphi - x_{\widehat{\varphi}_p}\|_2^q &\leq \langle x_\varphi - x_{\widehat{\varphi}_p}, \nabla_x \widehat{\varphi}_p(x_\varphi; w) - \nabla \varphi(x_\varphi) \rangle + \tfrac{\mathrm{Lip}(\nabla^p \varphi)}{p!} \langle (x_\varphi - x_{\widehat{\varphi}_p}) \|x_\varphi - w\|_2^{p-1}, x_\varphi - w \rangle \\ &= f(x_\varphi) - \textstyle\sum_{i=0}^{p-1} \frac{1}{i!} \nabla^i f(w)[x_\varphi - w]^{\otimes i} + \tfrac{\mathrm{Lip}(\nabla^p \varphi)}{p!} \langle (x_\varphi - x_{\widehat{\varphi}_p}) \|x_\varphi - w\|_2^{p-1}, (x_\varphi - w) \rangle \\ &\leq \tfrac{\mathrm{Lip}(\nabla^{p-1} f)}{p!} \|x_\varphi - w\|_2^p + \|x_\varphi - x_{\widehat{\varphi}_p}\|_2 \tfrac{\mathrm{Lip}(\nabla^p \varphi)}{p!} \|x_\varphi - w\|_2^p \leq \|x_\varphi - x_{\widehat{\varphi}_p}\|_2 \tfrac{2\,\mathrm{Lip}(\nabla^p \varphi)}{p!} \|x_\varphi - w\|_2^p.\end{aligned}$$

□

Fix any $\lambda \in \Lambda$ and $i \in [n]$, and consider the choices $\varphi = m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$, $\varphi_0 \equiv 0$, and $w = \hat{\beta}(\lambda)$. By Assump. 1f, $\widehat{\varphi}_p(\cdot; w) + \frac{\mathrm{Lip}(\nabla^p \varphi)}{(p+1)!} \| \cdot - w\|_2^{p+1} + \varphi_0$ has $\nu(r) = \mu r^q$ gradient growth for $\mu = c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]$. Since $\mathrm{Lip}(\nabla^p \varphi) \leq C_{\ell,p+1} + \lambda C_{\pi,p+1}$ by Assump. 3b, the desired result (16b) follows from Lemma 18.

B.1.4 Proof of (17a): Proximity of $\mathbf{ACV}_p^{\mathbf{HO}}$ and CV

Fix any $\lambda \in \Lambda$. To control the discrepancy between $\mathbf{ACV}_p^{\mathrm{HO}}(\lambda)$ and $\mathbf{CV}(\lambda)$, we first rewrite the difference using Taylor's theorem with Lagrange remainder:

$$\begin{aligned}
\mathbf{ACV}_p^{\mathrm{HO}}(\lambda) - \mathbf{CV}(\lambda) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(z_i, \tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda)) - \ell(z_i, \hat{\beta}_{\text{-}i}(\lambda)) \\
&= \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}_{\text{-}i}(\lambda)), \tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\rangle + \tfrac{1}{2}\nabla_\beta^2 \ell(z_i, \tilde{s}_i)[\tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)]^{\otimes 2}
\end{aligned}$$

for some $\tilde{s}_i \in \{t\tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p} + (1-t)\hat{\beta}_{\text{-}i}(\lambda) : t \in [0,1]\}$. We next use the mean-value theorem to expand each function $\langle \nabla_\beta \ell(z_i, \cdot), \tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\rangle$ around the full-data estimator $\hat{\beta}(\lambda)$:

$$\begin{aligned}
\mathbf{ACV}_p^{\mathrm{HO}}(\lambda) - \mathbf{CV}(\lambda) = \tfrac{1}{n}\textstyle\sum_{i=1}^n &\langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda)), \tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\rangle + \tfrac{1}{2}\nabla_\beta^2 \ell(z_i, \tilde{s}_i)[\tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)]^{\otimes 2} \\
&+ \langle \nabla_\beta^2 \ell(z_i, s_i)(\hat{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)), \tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\rangle
\end{aligned}$$

for some $s_i \in \{t\hat{\beta}(\lambda) + (1-t)\hat{\beta}_{\text{-}i}(\lambda) : t \in [0,1]\}$. Finally, we invoke Cauchy-Schwarz, the definition of the operator norm, the estimator proximity results (15) and (16a), and Assump. 2 to obtain

$$\begin{aligned}
|\mathbf{ACV}_p^{\mathrm{HO}}(\lambda) - \mathbf{CV}(\lambda)| &\le \tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 \|\tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2 + \tfrac{1}{2}\|\nabla_\beta^2 \ell(z_i, \tilde{s}_i)\|_{\mathrm{op}} \|\tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2^2 \\
&\quad + \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\mathrm{op}} \|\hat{\beta}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2 \|\tilde{\beta}_{\text{-}i}^{\mathrm{HO}_p}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2 \\
&\le \tfrac{1}{n}\textstyle\sum_{i=1}^n (\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 \|\hat{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2^{\frac{p}{q-1}} \\
&\quad + \tfrac{1}{2}(\kappa_{p,\lambda}^\lambda)^{\frac{2}{q-1}} \|\nabla_\beta^2 \ell(z_i, \tilde{s}_i)\|_{\mathrm{op}} \|\hat{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2^{\frac{2p}{q-1}} \\
&\quad + (\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}} \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\mathrm{op}} \|\hat{\beta}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2^{\frac{p+q-1}{q-1}} \\
&\le \frac{1}{n^{\frac{p}{(q-1)^2}}} \frac{(\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}}}{c_m^{\frac{p}{(q-1)^2}}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^{\frac{p+(q-1)^2}{(q-1)^2}} \\
&\quad + \tfrac{1}{2} \frac{1}{n^{\frac{2p}{(q-1)^2}}} \frac{(\kappa_{p,\lambda}^\lambda)^{\frac{2}{q-1}}}{c_m^{\frac{2p}{(q-1)^2}}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, \tilde{s}_i)\|_{\mathrm{op}} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^{\frac{2p}{(q-1)^2}} \\
&\quad + \frac{1}{n^{\frac{p+q-1}{(q-1)^2}}} \frac{(\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}}}{c_m^{\frac{p+q-1}{(q-1)^2}}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\mathrm{op}} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^{\frac{p+q-1}{(q-1)^2}} \\
&\quad \le \frac{1}{n^{\frac{p}{(q-1)^2}}} \frac{(\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}}}{c_m^{\frac{p}{(q-1)^2}}} \mathrm{B}^\ell_{0,\frac{p+(q-1)^2}{(q-1)^2}} + \frac{1}{2} \frac{1}{n^{\frac{2p}{(q-1)^2}}} \frac{(\kappa_{p,\lambda}^\lambda)^{\frac{2}{q-1}}}{c_m^{\frac{2p}{(q-1)^2}}} \mathrm{B}^\ell_{1,\frac{2p}{(q-1)^2}} + \frac{1}{n^{\frac{p+q-1}{(q-1)^2}}} \frac{(\kappa_{p,\lambda}^\lambda)^{\frac{1}{q-1}}}{c_m^{\frac{p+q-1}{(q-1)^2}}} \mathrm{B}^\ell_{1,\frac{p+q-1}{(q-1)^2}}.
\end{aligned}$$

B.1.5 Proof of (17b): Proximity of $\mathbf{ACV}_p$ and CV

The proof of the bound (17b) is identical to that of the bound (17a) once we substitute $2\kappa_{p,\lambda}^\lambda$ for $\kappa_{p,\lambda}^\lambda$ by invoking (16b) in place of (16a).

C Proof of Prop. 3: Sufficient conditions for assumptions

We prove each of the independent claims in turn.

Assump. 3 holds This first claim follows from the triangle inequality and the definition of the Lipschitz constant Lip.

$\boldsymbol{\hat{\beta}(\lambda) \to \hat{\beta}(\infty)}$ For each $\lambda \in [0, \infty)$, by the Optimizer Comparison Lemma 1 with $\varphi_2 = \pi$ and $\varphi_1 = \frac{1}{\lambda} m(\mathbb{P}_n, \cdot, \lambda)$ and the nonnegativity of ℓ,

$$\nu_\pi(\|\hat{\beta}(\lambda) - \hat{\beta}(\infty)\|_2) \le \tfrac{1}{\lambda}(\ell(\mathbb{P}_n, \hat{\beta}(\infty)) - \ell(\mathbb{P}_n, \hat{\beta}(\lambda))) \le \tfrac{1}{\lambda}\ell(\mathbb{P}_n, \hat{\beta}(\infty)).$$

Therefore, $\nu_\pi(\|\hat{\beta}(\lambda) - \hat{\beta}(\infty)\|_2) \to 0$ as $\lambda \to \infty$. Now, since ν_π is increasing, its inverse ω_π is increasing with $\omega_\pi(0) = 0$, and hence we have $\|\hat{\beta}(\lambda) - \hat{\beta}(\infty)\|_2 \to 0$ as $\lambda \to \infty$.

Assump. 1 holds Fix any $\Lambda \subseteq [0, \infty]$, and let mineig denote the minimum eigenvalue. The local strong convexity of π implies that there exist a neighborhood $\mathcal{N}$ of $\hat{\beta}(\infty)$ and some $c_\pi > 0$ for which $\nabla^2 \pi(\beta) \succeq c_\pi \mathrm{Id}$ for all $\beta \in \mathcal{N}$. Since $\hat{\beta}(\lambda) \to \hat{\beta}(\infty)$ as $\lambda \to \infty$, there exists $\lambda_\pi < \infty$ such that $\hat{\beta}(\lambda) \in \mathcal{N}$ for all $\lambda \geq \lambda_\pi$. Hence, for any $\lambda, \lambda' \in \Lambda$ and $i \in [n]$, we may use the c_m-strong convexity of $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda')$ and $m(\mathbb{P}_{n,\text{-}i}, \cdot, 0) = \ell(\mathbb{P}_{n,\text{-}i}, \cdot)$ to conclude that

$$\mathrm{mineig}(\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda')) = \mathrm{mineig}(\nabla_\beta^2 \ell(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda)) + \lambda' \nabla_\beta^2 \pi(\hat{\beta}(\lambda))) \geq \max(c_m, (c_m + \lambda' c_\pi)\mathbb{I}[\lambda \geq \lambda_\pi]).$$

Furthermore, the c_m-strong convexity and differentiability of $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$ imply that $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$ has $\nu_m(r) = c_m r^2$ gradient growth. Thus, Assump. 1 is satisfied for Λ.

Assump. 2 holds Fix any $\Lambda \subseteq [0, \infty]$ and $\lambda \in \Lambda$. For each $i \in [n]$, the triangle inequality and the definition of the Lipschitz constant imply

$$\begin{aligned}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 &\leq \|\nabla_\beta \ell(z_i, \hat{\beta}(\infty))\|_2 + \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda)) - \nabla_\beta \ell(z_i, \hat{\beta}(\infty))\|_2 \\ &\leq \|\nabla_\beta \ell(z_i, \hat{\beta}(\infty))\|_2 + L_i \|\hat{\beta}(\lambda) - \hat{\beta}(\infty)\|_2.\end{aligned}$$

Moreover, since $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$ is c_m-strongly convex and the minimum eigenvalue is a concave function, Jensen's inequality gives for each β

$$\mathrm{mineig}(m(\mathbb{P}_n, \beta, \lambda)) = \mathrm{mineig}(\tfrac{1}{n-1} \textstyle\sum_{i=1}^n m(\mathbb{P}_{n,\text{-}i}, \beta, \lambda)) \geq \tfrac{1}{n-1} \textstyle\sum_{i=1}^n \mathrm{mineig}(m(\mathbb{P}_{n,\text{-}i}, \beta, \lambda)) \geq \tfrac{n}{n-1} c_m.$$

Hence $m(\mathbb{P}_n, \cdot, \lambda)$ has $\nu_m(r) = \frac{n}{n-1} c_m r^2$ gradient growth, and the Optimizer Comparison Lemma 1 with $\varphi_2 = \lambda\pi$ and $\varphi_1 = m(\mathbb{P}_n, \cdot, \lambda)$ and Cauchy-Schwarz imply

$$\tfrac{n}{n-1} c_m \|\hat{\beta}(\lambda) - \hat{\beta}(\infty)\|_2 \leq \|\nabla_\beta \ell(\mathbb{P}_n, \hat{\beta}(\infty))\|_2.$$

Therefore,

$$\mathrm{B}_{s,r}^\ell \leq \tfrac{1}{n} \textstyle\sum_{i=1}^n L_i^s (\|\nabla_\beta \ell(z_i, \hat{\beta}(\infty))\|_2 + \tfrac{n-1}{n} \tfrac{L_i}{c_m} \|\nabla_\beta \ell(\mathbb{P}_n, \hat{\beta}(\infty))\|_2)^r < \infty.$$

D Proof of Thm. 4: ACV$^{\mathrm{IJ}}$-ACV assessment error

We will prove the following more detailed statement from which Thm. 4 immediately follows.

Theorem 19 (**ACV$^{\mathrm{IJ}}$-ACV** assessment error)**.** *If Assump. 1 holds for $\Lambda \subseteq [0, \infty]$, then, for each $\lambda \in \Lambda$,*

$$\|\tilde{\beta}_{\text{-}i}^{IJ}(\lambda) - \tilde{\beta}_{\text{-}i}(\lambda)\|_2 \leq \frac{\|\nabla_\beta^2 \ell(z_i, \hat{\beta}(\lambda))\|_{\mathrm{op}} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2}{c_{\lambda,\lambda}^2 n^2} \tag{19}$$

where $c_{\lambda,\lambda} \triangleq c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]$. If, in addition, Assump. 2 holds for Λ and each $(s, r) \in \{(1,2), (2,2), (3,2)\}$, then

$$|\mathbf{ACV}^{\mathrm{IJ}}(\lambda) - \mathbf{ACV}(\lambda)| \leq \frac{\mathrm{B}_{1,2}^\ell}{c_{\lambda,\lambda}^2 n^2} + \frac{\mathrm{B}_{2,2}^\ell}{c_{\lambda,\lambda}^3 n^3} + \frac{\mathrm{B}_{3,2}^\ell}{2 c_{\lambda,\lambda}^4 n^4}. \tag{20}$$

D.1 Proof of (19): Proximity of ACV and ACV$^{\mathrm{IJ}}$ estimators

We begin with a lemma that controls the discrepancy between two Newton (or, more generally, proximal Newton) estimators. Recall the definition of the proximal operator $\mathrm{prox}_H^{\varphi_0}$ (11).

Lemma 20 (Proximal Newton comparison)**.** *For any $\beta, g \in \mathbb{R}^d$, invertible $H, \tilde{H} \in \mathbb{R}^{d \times d}$, and convex φ_0, the proximal Newton estimators*

$$\beta_H = \mathrm{prox}_H^{\varphi_0}(\beta - H^{-1} g) \quad \textit{and} \quad \beta_{\tilde{H}} = \mathrm{prox}_{\tilde{H}}^{\varphi_0}(\beta - \tilde{H}^{-1} g)$$

satisfy

$$\|\beta_H - \beta_{\tilde{H}}\|_2 \leq \frac{\|(\tilde{H} - H)(\beta_H - \beta)\|_2}{\mathrm{mineig}(\tilde{H}) \vee 0} \leq \frac{\|\tilde{H} - H\|_{\mathrm{op}} \|\beta_H - \beta\|_2}{\mathrm{mineig}(\tilde{H}) \vee 0}.$$

Proof If $\mathrm{mineig}(\tilde{H}) \leq 0$, the claim is vacuous, so assume $\mathrm{mineig}(\tilde{H}) > 0$. Writing $\varphi_2(x) = \frac{1}{2}\|\beta - \tilde{H}^{-1}g - x\|_{\tilde{H}}^2 + \varphi_0(x)$ and $\varphi_1(x) = \frac{1}{2}\|\beta - H^{-1}g - x\|_H^2 + \varphi_0(x)$, note that $\beta_{\tilde{H}} = \mathrm{argmin}_x\, \varphi_2(x)$ and $\beta_H = \mathrm{argmin}_x\, \varphi_1(x)$ by the definition of the proximal operator (11). Importantly, φ_2 is subdifferentiable and satisfies the gradient growth property with $\nu_{\varphi_2}(r) = \mathrm{mineig}(\tilde{H})r^2$. Invoking the Optimizer Comparison Lemma 1 and Cauchy-Schwarz, we have

$$\mathrm{mineig}(\tilde{H})\|\beta_H - \beta_{\tilde{H}}\|_2^2 \leq \langle \tilde{H}(\beta - \beta_H) - g - H(\beta - \beta_H) + g, \beta_H - \beta_{\tilde{H}}\rangle \leq \|(\tilde{H} - H)(\beta - \beta_H)\|_2\|\beta_H - \beta_{\tilde{H}}\|_2.$$

Rearranging both sides gives the first advertised inequality. □

Now fix any $\lambda \in \Lambda$ and $i \in [n]$, and let

$$\tilde{H} = \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda) \quad \text{and} \quad H = \nabla_\beta^2 m(\mathbb{P}_n, \hat{\beta}(\lambda), \lambda) = \tfrac{n}{n-1}\tfrac{1}{n}\textstyle\sum_{j=1}^n \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}j}, \hat{\beta}(\lambda), \lambda).$$

By Assump. 1, $\mathrm{mineig}(\tilde{H}) \geq c_{\lambda,\lambda}$. Moreover, Assump. 1, the concavity of the minimum eigenvalue, and Jensen's inequality imply

$$\mathrm{mineig}(H) \geq \tfrac{n}{n-1}\tfrac{1}{n}\textstyle\sum_{j=1}^n \mathrm{mineig}(\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}j}, \hat{\beta}(\lambda), \lambda)) \geq \tfrac{n}{n-1}c_{\lambda,\lambda} \geq c_{\lambda,\lambda}.$$

Hence, we may apply Lemma 20 with $\beta_H = \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda)$, $\beta_{\tilde{H}} = \tilde{\beta}_{\text{-}i}(\lambda)$, $\beta = \hat{\beta}(\lambda)$, and $\varphi_0 \equiv 0$ to find that

$$\begin{aligned}\|\tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda) - \tilde{\beta}_{\text{-}i}(\lambda)\|_2 &\leq \tfrac{1}{c_{\lambda,\lambda}}\|\nabla_\beta^2 m(\mathbb{P}_n, \hat{\beta}(\lambda), \lambda) - \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)\|_{\mathrm{op}}\|\tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda) - \hat{\beta}(\lambda)\|_2 \\ &= \tfrac{1}{n^2}\tfrac{1}{c_{\lambda,\lambda}}\|\nabla_\beta^2 \ell(z_i, \hat{\beta}(\lambda))\|_{\mathrm{op}}\|\nabla_\beta^2 m(\mathbb{P}_n, \hat{\beta}(\lambda), \lambda)^{-1}\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 \\ &\leq \tfrac{1}{n^2}\tfrac{1}{c_{\lambda,\lambda}^2}\|\nabla_\beta^2 \ell(z_i, \hat{\beta}(\lambda))\|_{\mathrm{op}}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2.\end{aligned}$$

D.2 Proof of (20): Proximity of ACV and $\mathbf{ACV}^{\mathrm{IJ}}$

Fix any $\lambda \in \Lambda$. To control the discrepancy between $\mathbf{ACV}(\lambda)$ and $\mathbf{ACV}^{\mathrm{IJ}}(\lambda)$, we first rewrite the difference using Taylor's theorem with Lagrange remainder:

$$\begin{aligned}\mathbf{ACV}^{\mathrm{IJ}}(\lambda) - \mathbf{ACV}(\lambda) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(z_i, \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda)) - \ell(z_i, \tilde{\beta}_{\text{-}i}(\lambda)) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \tilde{\beta}_{\text{-}i}(\lambda)), \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda) - \tilde{\beta}_{\text{-}i}(\lambda)\rangle + \tfrac{1}{2}\nabla_\beta^2 \ell(z_i, \tilde{s}_i)[\tilde{\beta}_{\text{-}i}(\lambda) - \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda)]^{\otimes 2}\end{aligned}$$

for some $\tilde{s}_i \in \{t\tilde{\beta}_{\text{-}i}(\lambda) + (1-t)\tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda) : t \in [0,1]\}$. We next use the mean-value theorem to expand each function $\langle \nabla_\beta \ell(z_i, \cdot), \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda) - \tilde{\beta}_{\text{-}i}(\lambda)\rangle$ around the full-data estimator $\hat{\beta}(\lambda)$:

$$\begin{aligned}\mathbf{ACV}^{\mathrm{IJ}}(\lambda) - \mathbf{ACV}(\lambda) = \tfrac{1}{n}\textstyle\sum_{i=1}^n &\langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda)), \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda) - \tilde{\beta}_{\text{-}i}(\lambda)\rangle + \tfrac{1}{2}\nabla_\beta^2 \ell(z_i, \tilde{s}_i)[\tilde{\beta}_{\text{-}i}(\lambda) - \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda)]^{\otimes 2} \\ &+ \langle \nabla_\beta^2 \ell(z_i, s_i)(\tilde{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)), \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda) - \tilde{\beta}_{\text{-}i}(\lambda)\rangle\end{aligned}$$

for some $s_i \in \{t\hat{\beta}(\lambda) + (1-t)\tilde{\beta}_{\text{-}i}(\lambda) : t \in [0,1]\}$. Now, by Assump. 1, we have

$$\|\tilde{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2 = \tfrac{1}{n}\|\mathrm{H}_i^{-1}\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 \leq \tfrac{1}{nc_{\lambda,\lambda}}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2.$$

Combining these observations with Cauchy-Schwarz, the definition of the operator norm, the estimator proximity result (19), the definition of the Lipschitz constant $\mathrm{Lip}(\nabla_\beta \ell(z_i, \cdot))$, and Assump. 2 we obtain

$$\begin{aligned}|\mathbf{ACV}^{\mathrm{IJ}}(\lambda) - \mathbf{ACV}(\lambda)| &\leq \tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2\|\tilde{\beta}_{\text{-}i}(\lambda) - \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda)\|_2 + \tfrac{1}{2}\|\nabla_\beta^2 \ell(z_i, \tilde{s}_i)\|_{\mathrm{op}}\|\tilde{\beta}_{\text{-}i}(\lambda) - \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda)\|_2^2 \\ &\quad + \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\mathrm{op}}\|\tilde{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2\|\tilde{\beta}_{\text{-}i}(\lambda) - \tilde{\beta}_{\text{-}i}^{\mathrm{IJ}}(\lambda)\|_2 \\ &\leq \tfrac{1}{n^2c_{\lambda,\lambda}^2}\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, \hat{\beta}(\lambda))\|_{\mathrm{op}}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^2 \\ &\quad + \tfrac{1}{2}\tfrac{1}{n^4c_{\lambda,\lambda}^4}\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, \tilde{s}_i)\|_{\mathrm{op}}\|\nabla_\beta^2 \ell(z_i, \hat{\beta}(\lambda))\|_{\mathrm{op}}^2\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^2 \\ &\quad + \tfrac{1}{n^3c_{\lambda,\lambda}^3}\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\mathrm{op}}\|\nabla_\beta^2 \ell(z_i, \hat{\beta}(\lambda))\|_{\mathrm{op}}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^2 \\ &\leq \tfrac{1}{n^2c_{\lambda,\lambda}^2}\mathrm{B}_{1,2}^{\ell} + \tfrac{1}{2n^4c_{\lambda,\lambda}^4}\mathrm{B}_{3,2}^{\ell} + \tfrac{1}{n^3c_{\lambda,\lambda}^3}\mathrm{B}_{2,2}^{\ell}.\end{aligned}$$

E Proof of Thm. 6: ACV-CV selection error

The first claim follows immediately from the following more detailed version of Thm. 6.

Theorem 21 (**ACV** proximity implies $\hat{\beta}$ proximity)**.** *Suppose Assumps. 1 and 3 hold for some* $\Lambda \subseteq [0,\infty]$ *and each* $(s,r) \in \{(0,2),(1,2)\}$*. Then, for all* $\lambda', \lambda \in \Lambda$ *with* $\lambda' < \lambda$,

$$\|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2^2 \leq C_{1,\lambda,\lambda'}\Big(\tfrac{\tilde{C}_{2,\lambda,\lambda'}}{n} + \mathbf{ACV}(\lambda) - \mathbf{ACV}(\lambda') + \tfrac{C_{3,\lambda,\lambda'}}{n^2}\Big),$$

for $C_{1,\lambda,\lambda'}$ *and* $C_{3,\lambda,\lambda'}$ *defined in Thm. 7 and*

$$\tilde{C}_{2,\lambda,\lambda'} = \tfrac{3\mathrm{B}_{0,2}^{\ell}}{c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]} + \tfrac{\mathrm{B}_{0,2}^{\ell}}{c_\ell + \lambda' c_\pi \mathbb{I}[\lambda' \geq \lambda_\pi]}.$$

Proof Fix any $\lambda', \lambda \in \Lambda$ with $\lambda' < \lambda$. We will proceed precisely an in the proof of Thm. 23, except we will provide alternative bounds for the quantities ΔT_2 and ΔT_3 in the loss decomposition (26). First, we apply Cauchy-Schwarz, the definition of the operator norm, the triangle inequality, and the arithmetic-geometric mean inequality in turn to find

$$\begin{aligned}
|\Delta T_2| &= \tfrac{1}{n}|\tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda)) - \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')), \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1} \nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\rangle| \\
&\leq \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1}\|_{\mathrm{op}} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 (\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 + \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2) \\
&\leq \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1}\|_{\mathrm{op}} (\tfrac{3}{2}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^2 + \tfrac{1}{2}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2^2) \\
&\leq \tfrac{1}{n}\tfrac{2\mathrm{B}_{0,2}^{\ell}}{c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]},
\end{aligned} \tag{21}$$

where we have used Assump. 1 and Assump. 2 for $(s,r) = (0,2)$ in the final line.

Next, we again apply the triangle inequality, the definition of the operator norm, the arithmetic-geometric mean inequality, Assump. 1, and Assump. 2 for $(s,r) = (0,2)$ to obtain

$$\begin{aligned}
|\Delta T_3| = &|\tfrac{1}{n^2}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')), \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1} \nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\rangle \\
&\quad - \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')), \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda'), \lambda')^{-1} \nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\rangle| \\
&\leq \tfrac{1}{n}\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2 \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1}\|_{\mathrm{op}} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 \\
&\quad + \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2^2 \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda'), \lambda')^{-1}\|_{\mathrm{op}} \\
&\leq \tfrac{1}{n}\tfrac{1}{n}\textstyle\sum_{i=1}^n (\tfrac{1}{2}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2^2 + \tfrac{1}{2}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^2) \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1}\|_{\mathrm{op}} \\
&\quad + \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2^2 \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda'), \lambda')^{-1}\|_{\mathrm{op}} \\
&\leq \tfrac{1}{n}(\tfrac{\mathrm{B}_{0,2}^{\ell}}{c_\ell + \lambda c_\pi \mathbb{I}[\lambda \geq \lambda_\pi]} + \tfrac{\mathrm{B}_{0,2}^{\ell}}{c_\ell + \lambda' c_\pi \mathbb{I}[\lambda' \geq \lambda_\pi]}).
\end{aligned} \tag{22}$$

Plugging the bounds (21) and (22) into the proof Thm. 23 yields the result. □
The second claim (10) follows the first and the following bound on $|\mathbf{ACV}(\lambda_{\mathbf{CV}}) - \mathbf{ACV}(\lambda_{\mathbf{ACV}})|$.

Lemma 22. *Suppose Assumps. 1, 2, and 3 hold for some* $\Lambda \subseteq [0,\infty]$ *and each* $(s,r) \in \{(0,3),(1,3),(1,4)\}$*. If* $\lambda_{\mathbf{ACV}} \in \operatorname{argmin}_{\lambda\in\Lambda} \mathbf{ACV}(\lambda)$ *and* $\lambda_{\mathbf{CV}} \in \operatorname{argmin}_{\lambda\in\Lambda} \mathbf{CV}(\lambda)$*, then*

$$0 \leq \mathbf{ACV}(\lambda_{\mathbf{CV}}) - \mathbf{ACV}(\lambda_{\mathbf{ACV}}) \leq 2(\tfrac{\kappa_2}{n^2}\tfrac{\mathrm{B}_{0,3}^{\ell}}{c_m^2} + \tfrac{\kappa_2}{n^3}\tfrac{\mathrm{B}_{1,3}^{\ell}}{c_m^3} + \tfrac{\kappa_2^2}{n^4}\tfrac{\mathrm{B}_{1,4}^{\ell}}{2c_m^4}).$$

Proof Since $\lambda_{\mathbf{CV}}$ minimizes $\mathbf{CV}$ and $\lambda_{\mathbf{ACV}}$ minimizes $\mathbf{ACV}$,

$$0 \leq \mathbf{ACV}(\lambda_{\mathbf{CV}}) - \mathbf{ACV}(\lambda_{\mathbf{ACV}}) \leq \mathbf{ACV}(\lambda_{\mathbf{CV}}) - \mathbf{ACV}(\lambda_{\mathbf{ACV}}) + \mathbf{CV}(\lambda_{\mathbf{ACV}}) - \mathbf{CV}(\lambda_{\mathbf{CV}}).$$

The result now follows from two applications of Thm. 2. □

F Proof of Thm. 7: Strong ACV-CV selection error

The first claim follows immediately from the following more detailed version of Thm. 7, proved in App. F.1.

Theorem 23 (Strong **ACV** proximity implies $\hat{\beta}$ proximity)**.** *Suppose Assumps. 1, 2, 3, and 4 hold for some* $\Lambda \subseteq [0, \infty]$ *with* $0 \in \Lambda$ *and each* $(s, r) \in \{(0, 2), (1, 1), (1, 2)\}$. *Suppose also* $\|\nabla\pi(\hat{\beta}(0))\|_2 > 0$. *Then for all* $\lambda', \lambda \in \Lambda$ *with* $\lambda' < \lambda$,

$$\|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2^2 \leq C_{1,\lambda,\lambda'}\Big(C_{2,\lambda,\lambda'}/n\|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2 + \mathbf{ACV}(\lambda) - \mathbf{ACV}(\lambda') + C_{3,\lambda,\lambda'}/n^2\Big) \quad \textit{and hence} \tag{23}$$

$$|\|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2 - \tfrac{C_{1,\lambda,\lambda'}C_{2,\lambda,\lambda'}}{2n}| \leq \sqrt{\tfrac{C_{1,\lambda,\lambda'}^2 C_{2,\lambda,\lambda'}^2}{4n^2} + C_{1,\lambda,\lambda'}(\mathbf{ACV}(\lambda) - \mathbf{ACV}(\lambda')) + \tfrac{C_{1,\lambda,\lambda'}C_{3,\lambda,\lambda'}}{n^2}}, \tag{24}$$

where

$$\begin{aligned}
C_{1,\lambda,\lambda'} &= \tfrac{2}{c_m}\tfrac{\lambda-\lambda'}{\lambda+\lambda'}\tfrac{n-1}{n}, \\
C_{2,\lambda,\lambda'} &= \tfrac{2\mathrm{B}_{1,1}^{\ell}}{c_\ell+\lambda c_\pi \mathbb{I}[\lambda\geq\lambda_\pi]} + \tfrac{2\mathrm{B}_{0,2}^{\ell}\kappa_{2,\lambda'}^{\lambda'}}{c_\ell+\lambda' c_\pi \mathbb{I}[\lambda\geq\lambda_\pi]} + \tfrac{n-1}{n}\tfrac{\mathrm{B}_{0,2}^{\ell}C_{\pi,2}\kappa_{1,\lambda}^{\lambda}\kappa_{1,\lambda'}^{\lambda}}{\|\nabla\pi(\hat{\beta}(0))\|_2 c_m}, \\
C_{3,\lambda,\lambda'} &= \tfrac{\mathrm{B}_{1,2}^{\ell}}{c_m^2}
\end{aligned}$$

for $\kappa_{p,\lambda'}^{\lambda}$ *defined in Thm. 14.*

The second claim follows directly from the Thm. 23 bound (24) and Lemma 22.

F.1 Proof of Thm. 23

Fix any $\lambda', \lambda \in \Lambda$ with $\lambda' < \lambda$. The statement (24) follows directly from (23) and the quadratic formula, so we will focus on establishing the bound (23). We begin by writing the difference in estimator training losses as a difference in **ACV** values plus a series of error terms:

$$\ell(\mathbb{P}_n, \hat{\beta}(\lambda)) - \ell(\mathbb{P}_n, \hat{\beta}(\lambda')) = \mathbf{ACV}(\lambda) - \mathbf{ACV}(\lambda') + \Delta T_1 - \Delta T_2 - \Delta T_3 \tag{25}$$

for

$$\begin{aligned}
\Delta T_1 &\triangleq \widehat{\mathbf{ACV}}(\lambda) - \mathbf{ACV}(\lambda) + \mathbf{ACV}(\lambda') - \widehat{\mathbf{ACV}}(\lambda'), \\
\widehat{\mathbf{ACV}}(\lambda) &\triangleq \ell(\mathbb{P}_n, \hat{\beta}(\lambda)) + \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda)), \tilde{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\rangle, \\
\Delta T_2 &\triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda)) - \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')), \tilde{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\rangle, \quad \text{and} \\
\Delta T_3 &\triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')), \tilde{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda) - (\tilde{\beta}_{\text{-}i}(\lambda') - \hat{\beta}(\lambda'))\rangle.
\end{aligned} \tag{26}$$

Here, $\widehat{\mathbf{ACV}}(\lambda)$ arises by first-order Taylor-expanding each $\ell(\mathbb{P}_n, \tilde{\beta}_{\text{-}i}(\lambda))$ about $\hat{\beta}(\lambda)$ in the expression of $\mathbf{ACV}(\lambda)$. To complete the proof, we will bound ΔT_1, ΔT_2, ΔT_3, and $\ell(\mathbb{P}_n, \hat{\beta}(\lambda)) - \ell(\mathbb{P}_n, \hat{\beta}(\lambda'))$ in turn.

F.1.1 Bounding ΔT_1

To control ΔT_1, we will appeal to the following lemma which shows that $\widehat{\mathbf{ACV}}$ provides an $O(1/n^2)$ approximation to **ACV**, uniformly in λ. The proof can be found App. F.2.

Lemma 24 (**ACV**-$\widehat{\mathbf{ACV}}$ approximation error)**.** *Suppose Assumps. 1 and 2 hold for some* $\Lambda \subseteq [0, \infty]$ *and* $(s, r) = (1, 2)$. *Then, for each* $\lambda \in \Lambda$,

$$|\mathbf{ACV}(\lambda) - \widehat{\mathbf{ACV}}(\lambda)| \leq \tfrac{1}{n^2}\tfrac{\mathrm{B}_{1,2}^{\ell}}{2c_m^2}.$$

Applying Lemma 24 to λ and λ', we obtain

$$|\Delta T_1| \leq \tfrac{1}{n^2}\tfrac{\mathrm{B}_{1,2}^{\ell}}{c_m^2}. \tag{27}$$

F.1.2 Bounding ΔT_2

We employ the mean value theorem, Cauchy-Schwarz, the definition of the operator norm, Assump. 1, and Assump. 2 for $(s, r) = (1, 1)$ to find that

$$\begin{aligned}
|\Delta T_2| &= \tfrac{1}{n}|\tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat\beta(\lambda)) - \nabla_\beta \ell(z_i, \hat\beta(\lambda')), \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat\beta(\lambda), \lambda)^{-1} \nabla_\beta \ell(z_i, \hat\beta(\lambda))\rangle| \\
&= \tfrac{1}{n}|\tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta^2 \ell(z_i, s_{\lambda,\lambda'})(\hat\beta(\lambda) - \hat\beta(\lambda')), \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat\beta(\lambda), \lambda)^{-1} \nabla_\beta \ell(z_i, \hat\beta(\lambda))\rangle| \\
&\le \tfrac{1}{n}\|\hat\beta(\lambda) - \hat\beta(\lambda')\|_2 \tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat\beta(\lambda), \lambda)^{-1}\|_{\text{op}} \|\nabla_\beta^2 \ell(z_i, s_{\lambda,\lambda'})\|_{\text{op}} \|\nabla_\beta \ell(z_i, \hat\beta(\lambda))\|_2 \\
&\le \tfrac{1}{n}\|\hat\beta(\lambda) - \hat\beta(\lambda')\|_2 \tfrac{\mathrm{B}_{1,1}^\ell}{c_\ell + \lambda c_\pi \mathbb{I}[\lambda \ge \lambda_\pi]} \qquad (28)
\end{aligned}$$

for some convex combination $s_{\lambda,\lambda'}$ of $\hat\beta(\lambda)$ and $\hat\beta(\lambda')$.

F.1.3 Bounding ΔT_3

We next show that the double difference term ΔT_3 is controlled by estimator proximity $\|\hat\beta(\lambda) - \hat\beta(\lambda')\|_2$ and regularization parameter proximity $|\lambda - \lambda'|$ times an extra factor of $1/n$. This result is proved in App. F.3.

Lemma 25. *Suppose Assumps. 1, 2, 3, and 4 hold for some $\Lambda \subseteq [0, \infty]$ and each $(s, r) = \{(0, 2), (1, 1)\}$. Then, for all $\lambda, \lambda' \in \Lambda$,*

$$\begin{aligned}
|\Delta T_3| \le \tfrac{1}{n}\|\hat\beta(\lambda) - \hat\beta(\lambda')\|_2 \Big(\tfrac{\mathrm{B}_{1,1}^\ell}{c_\ell + \lambda c_\pi \mathbb{I}[\lambda \ge \lambda_\pi]} + \tfrac{2\mathrm{B}_{0,2}^\ell \kappa_{2,\lambda'}^{\lambda'}}{c_\ell + \lambda' c_\pi \mathbb{I}[\lambda \ge \lambda_\pi]}\Big) \qquad (29) \\
+ \tfrac{1}{n}|\lambda - \lambda'| \tfrac{\mathrm{B}_{0,2}^\ell C_{\pi,2}}{(c_\ell + c_\pi \lambda \mathbb{I}[\lambda \ge \lambda_\pi])(c_\ell + c_\pi \lambda' \mathbb{I}[\lambda \ge \lambda_\pi])}
\end{aligned}$$

for ΔT_3 defined in (26) and $\kappa_{2,\lambda'}^{\lambda'}$ defined in Thm. 14.

We combine (29) with the following bound on $|\lambda - \lambda'|$ proved in App. F.4:

Lemma 26 ($\hat\beta$ proximity implies λ proximity)**.** *Suppose Assumps. 1 and 4 hold for some $\Lambda \subseteq [0, \infty]$ with $0 \in \Lambda$. Then, for all $\lambda, \lambda' \in \Lambda$,*

$$|\lambda - \lambda'| \le \tfrac{n-1}{n} \tfrac{(C_{\ell,2} + \lambda C_{\pi,2})(C_{\ell,2} + \lambda' C_{\pi,2})}{c_m} \tfrac{1}{\|\nabla \pi(\hat\beta(0))\|_2} \|\hat\beta(\lambda) - \hat\beta(\lambda')\|_2. \qquad (30)$$

Together, (29) and (30) imply

$$|\Delta T_3| \le \tfrac{1}{n}\|\hat\beta(\lambda) - \hat\beta(\lambda')\|_2 \Big(\tfrac{\mathrm{B}_{1,1}^\ell}{c_\ell + \lambda c_\pi \mathbb{I}[\lambda \ge \lambda_\pi]} + \tfrac{2\mathrm{B}_{0,2}^\ell \kappa_{2,\lambda'}^{\lambda'}}{c_\ell + \lambda' c_\pi \mathbb{I}[\lambda \ge \lambda_\pi]} + \tfrac{n-1}{n} \tfrac{\mathrm{B}_{0,2}^\ell C_{\pi,2} \kappa_{1,\lambda}^\lambda \kappa_{1,\lambda'}^\lambda}{\|\nabla \pi(\hat\beta(0))\|_2 c_m}\Big). \qquad (31)$$

F.1.4 Putting the pieces together

Our final lemma, proved in App. F.5, establishes that, due to the curvature of the loss, two estimators with similar training loss must also be close in Euclidean norm.

Lemma 27 (Loss curvature)**.** *Suppose that for some $c_m > 0$ and $0 \le \lambda' < \lambda \le \infty$ and all $i \in [n]$, $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$ and $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda')$ have $\nu_m(r) = c_m r^2$ gradient growth. Then*

$$\tfrac{c_m}{2} \|\hat\beta(\lambda) - \hat\beta(\lambda')\|_2^2 \tfrac{\lambda + \lambda'}{\lambda - \lambda'} \le \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(\mathbb{P}_{n,\text{-}i}, \hat\beta(\lambda)) - \ell(\mathbb{P}_{n,\text{-}i}, \hat\beta(\lambda')) = \tfrac{n-1}{n}(\ell(\mathbb{P}_n, \hat\beta(\lambda)) - \ell(\mathbb{P}_n, \hat\beta(\lambda'))).$$

The advertised result (23) now follows by combining Lemma 27 with the loss difference decomposition (25) and the component bounds (27), (28), and (31).

F.2 Proof of Lemma 24: ACV-$\widehat{\textbf{ACV}}$ approximation error

By Taylor's theorem with Lagrange remainder,

$$\mathbf{ACV}(\lambda) - \widehat{\mathbf{ACV}}(\lambda) = \tfrac{1}{2n}\textstyle\sum_{i=1}^n \nabla_\beta^2 \ell(z_i, s_i)[\tilde\beta_{\text{-}i}(\lambda) - \hat\beta(\lambda)]^{\otimes 2}$$

for some $s_i \in \mathcal{L}_i = \{t\hat{\beta}(\lambda) + (1-t)\tilde{\beta}_{\text{-}i}(\lambda) : t \in [0,1]\}$. Assump. 1 implies that $m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)$ is c_m strongly convex and hence that $\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda) \succcurlyeq c_m \mathrm{I}_d$. Therefore, we may apply the definition of the operator norm and Assump. 2 to find that

$$\begin{aligned}
|\mathbf{ACV}(\lambda) - \widehat{\mathbf{ACV}}(\lambda)| &\le \tfrac{1}{2n} \textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\text{op}} \|\tilde{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2^2 \\
&= \tfrac{1}{2n} \textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\text{op}} \tfrac{1}{n^2} \|m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1} \nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^2 \\
&\le \tfrac{1}{2n} \textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\text{op}} \tfrac{1}{n^2} \tfrac{1}{c_m^2} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^2 \\
&\le \tfrac{1}{n^2} \tfrac{\mathrm{B}_{1,2}^\ell}{2c_m^2}.
\end{aligned}$$

F.3 Proof of Lemma 25: ΔT_3-bound

Fix any $\lambda, \lambda' \in \Lambda$. We first expand ΔT_3 into three terms:

$$\begin{aligned}
\Delta T_3 &= \tfrac{1}{n^2} \textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')), \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1} (\nabla_\beta \ell(z_i, \hat{\beta}(\lambda)) - \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')))\rangle \\
&\quad + \tfrac{1}{n^2} \textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')), (\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda')^{-1} - \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda'), \lambda')^{-1}) \nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\rangle \\
&\quad + \tfrac{1}{n^2} \textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')), (\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1} - \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda')^{-1}) \nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\rangle \\
&\quad \triangleq \Delta T_{31} + \Delta T_{32} + \Delta T_{33}.
\end{aligned}$$

Precisely as in (28) we obtain

$$|\Delta T_{31}| \le \tfrac{1}{n} \|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2 \tfrac{\mathrm{B}_{1,1}^\ell}{c_\ell + \lambda c_\pi \mathbb{I}[\lambda \ge \lambda_\pi]}.$$

Furthermore, we may use Cauchy-Schwarz, the definition of the operator norm, Assumps. 1 and 3, and Assump. 2 with $(s,r) = (0,2)$ to find

$$\begin{aligned}
|\Delta T_{32}| &\le \tfrac{1}{n}\tfrac{1}{n} \textstyle\sum_{i=1}^n \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2^2 \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda')^{-1}\|_{\text{op}} \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda'), \lambda')^{-1}\|_{\text{op}} \\
&\qquad \cdot \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda') - \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda'), \lambda')\|_{\text{op}} \\
&\le \tfrac{\mathrm{B}_{0,2}^\ell (C_{\ell,3} + \lambda' C_{\pi,3})}{(c_\ell + \lambda' c_\pi \mathbb{I}[\lambda \ge \lambda_\pi])(c_\ell + \lambda' c_\pi \mathbb{I}[\lambda' \ge \lambda_\pi])} \tfrac{1}{n} \|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2.
\end{aligned}$$

Finally, Cauchy-Schwarz, the definition of the operator norm, Assumps. 1 and 4, and Assump. 2 with $(s,r) = (0,2)$ yield

$$\begin{aligned}
|\Delta T_{33}| &\le \tfrac{1}{n}\tfrac{1}{n} \textstyle\sum_{i=1}^n \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1}\|_2 \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2^2 \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda') - \nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)\|_{\text{op}} \\
&\quad \cdot \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda')^{-1}\|_2 \\
&= \tfrac{1}{n} |\lambda - \lambda'| \|\nabla_\beta^2 \pi(\hat{\beta}(\lambda))\|_{\text{op}} \tfrac{1}{n} \textstyle\sum_{i=1}^n \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda)^{-1}\|_2 \|\nabla_\beta^2 m(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda), \lambda')^{-1}\|_2 \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2^2 \\
&\le \tfrac{1}{n} |\lambda - \lambda'| \tfrac{\mathrm{B}_{0,2}^\ell C_{\pi,2}}{(c_\ell + c_\pi \lambda \mathbb{I}[\lambda \ge \lambda_\pi])(c_\ell + c_\pi \lambda' \mathbb{I}[\lambda \ge \lambda_\pi])}.
\end{aligned}$$

We obtain the desired result by applying the triangle inequality and summing these three estimates.

F.4 Proof of Lemma 26: $|\lambda - \lambda'|$-bound

We begin with a lemma that allows us to rewrite a regularization parameter difference in terms of an estimator difference.

Lemma 28. *Fix any $\lambda, \lambda' \in [0, \infty]$. If $\nabla_\beta m(\mathbb{P}_n, \cdot, \lambda')$ is absolutely continuous, then*

$$\begin{aligned}
0 &= (\lambda - \lambda') \nabla_\beta \pi(\hat{\beta}(\lambda)) + \mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{\lambda,\lambda'}, \lambda')](\hat{\beta}(\lambda) - \hat{\beta}(\lambda')) \\
&= \tfrac{\lambda' - \lambda}{\lambda} \nabla_\beta \ell(\mathbb{P}_n, \hat{\beta}(\lambda)) + \mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{\lambda,\lambda'}, \lambda')](\hat{\beta}(\lambda) - \hat{\beta}(\lambda')).
\end{aligned}$$

for $S_{\lambda,\lambda'}$ distributed uniformly on the set $\{t\hat{\beta}(\lambda) + (1-t)\hat{\beta}(\lambda') : t \in [0,1]\}$.

Proof The first order optimality conditions for $\hat{\beta}(\lambda)$ and $\hat{\beta}(\lambda')$ and the absolute continuity of $\nabla_\beta m(\mathbb{P}_n, \cdot, \lambda')$ imply that

$$\begin{aligned}
0 &= \nabla_\beta m(\mathbb{P}_n, \hat{\beta}(\lambda), \lambda) = \nabla_\beta \ell(\mathbb{P}_n, \hat{\beta}(\lambda)) + \lambda \nabla \pi(\hat{\beta}(\lambda)) \\
&= (\lambda - \lambda')\nabla_\beta \pi(\hat{\beta}(\lambda)) + \nabla_\beta m(\mathbb{P}_n, \hat{\beta}(\lambda), \lambda') \\
&= (\lambda - \lambda')\nabla_\beta \pi(\hat{\beta}(\lambda)) + (\nabla_\beta m(\mathbb{P}_n, \hat{\beta}(\lambda), \lambda') - \nabla_\beta m(\mathbb{P}_n, \hat{\beta}(\lambda'), \lambda')) \\
&= (\lambda - \lambda')\nabla_\beta \pi(\hat{\beta}(\lambda)) + \mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{\lambda,\lambda'}, \lambda')](\hat{\beta}(\lambda) - \hat{\beta}(\lambda')) \\
&= \tfrac{\lambda' - \lambda}{\lambda} \nabla_\beta \ell(\mathbb{P}_n, \hat{\beta}(\lambda)) + \mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{\lambda,\lambda'}, \lambda')](\hat{\beta}(\lambda) - \hat{\beta}(\lambda'))
\end{aligned}$$

by Taylor's theorem with integral remainder. □

Now fix any $\lambda, \lambda' \in \Lambda$. Since $\nabla_\beta m(\mathbb{P}_n, \cdot, \lambda')$, $\nabla_\beta m(\mathbb{P}_n, \cdot, \lambda)$, and $\nabla_\beta m(\mathbb{P}_n, \cdot, 0) = \nabla_\beta \ell(\mathbb{P}_n, \cdot)$ are absolutely continuous by Assump. 4, we may apply Lemma 28 first to (λ, λ'), then to $(\lambda, 0)$, and finally to $(0, \lambda)$ to obtain

$$\begin{aligned}
0 &= \tfrac{\lambda' - \lambda}{\lambda} \nabla_\beta \ell(\mathbb{P}_n, \hat{\beta}(\lambda)) + \mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{\lambda,\lambda'}, \lambda')](\hat{\beta}(\lambda) - \hat{\beta}(\lambda')) \\
&= \tfrac{\lambda' - \lambda}{\lambda} \mathbb{E}[\nabla_\beta^2 \ell(\mathbb{P}_n, S_{\lambda,0})](\hat{\beta}(\lambda) - \hat{\beta}(0)) + \mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{\lambda,\lambda'}, \lambda')](\hat{\beta}(\lambda) - \hat{\beta}(\lambda')) \\
&= (\lambda - \lambda')\mathbb{E}[\nabla_\beta^2 \ell(\mathbb{P}_n, S_{\lambda,0})]\mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{0,\lambda}, \lambda)]^{-1} \nabla_\beta \pi(\hat{\beta}(0)) + \mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{\lambda,\lambda'}, \lambda')](\hat{\beta}(\lambda) - \hat{\beta}(\lambda'))
\end{aligned}$$

where $S_{\lambda,\lambda'}$ is distributed uniformly on the set $\{t\hat{\beta}(\lambda) + (1-t)\hat{\beta}(\lambda') : t \in [0,1]\}$ and $S_{\lambda,0}, S_{0,\lambda}$ are distributed uniformly on the set $\{t\hat{\beta}(\lambda) + (1-t)\hat{\beta}(0) : t \in [0,1]\}$. Rearranging and taking norms gives the identity

$$|\lambda - \lambda'| \|\nabla_\beta \pi(\hat{\beta}(0))\|_2 = \|\mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{\lambda,0}, \lambda)]\mathbb{E}[\nabla_\beta^2 \ell(\mathbb{P}_n, S_{\lambda,0})]^{-1}\mathbb{E}[\nabla_\beta^2 m(\mathbb{P}_n, S_{\lambda,\lambda'}, \lambda')](\hat{\beta}(\lambda) - \hat{\beta}(\lambda'))\|_2.$$

Our gradient growth assumption for the regularization parameter 0 implies that each $\ell(\mathbb{P}_{n,\text{-}i}, \cdot)$ is c_m-strongly convex [Nesterov, 2008, Lem. 1]. Therefore,

$$\mathbb{E}[\nabla_\beta^2 \ell(\mathbb{P}_n, S_{\lambda,0})] = \tfrac{n}{n-1}\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}[\nabla_\beta^2 \ell(\mathbb{P}_{n,\text{-}i}, S_{\lambda,0})] \succcurlyeq c_m \tfrac{n}{n-1} \mathrm{I}_d.$$

Applying this inequality along with Cauchy Schwarz and Assump. 4 for λ and λ', we now conclude that

$$|\lambda - \lambda'| \leq \tfrac{n-1}{n} \tfrac{(C_{\ell,2} + \lambda C_{\pi,2})(C_{\ell,2} + \lambda' C_{\pi,2})}{c_m} \tfrac{1}{\|\nabla \pi(\hat{\beta}(0))\|_2} \|\hat{\beta}(\lambda) - \hat{\beta}(\lambda')\|_2.$$

F.5 Proof of Lemma 27: Loss curvature

Fix any $\lambda, \lambda' \in \Lambda$ with $\lambda > \lambda'$ and $i \in [n]$, and consider the functions $\varphi_2 = m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda')$ and $\varphi_1 = \frac{\lambda'}{\lambda} m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$. The gradient growth condition in Assump. 1 implies that $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda')$ and $m(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda)$ are c_m-strongly convex. Hence, φ_2 admits a $\nu_{\varphi_2}(r) = \frac{c_m}{2} r^2$ error bound, and φ_1 admits a $\nu_{\varphi_1}(r) = \frac{\lambda'}{\lambda} \frac{c_m}{2} r^2$ error bound. The result now follows immediately from the optimizer comparison bound (4).

G Proof of Prop. 8

We write $\mathbb{E}_n[z_i] \triangleq \frac{1}{n}\sum_{i=1}^n z_i$ to denote a sample average. Consider the Lasso estimator

$$\hat{\beta}(\lambda) \triangleq \operatorname{argmin}_\beta \tfrac{1}{2n} \textstyle\sum_{i=1}^n (\beta - z_i)^2 + \lambda|\beta| = \max(\bar{z} - \lambda, 0).$$

Define $\epsilon_i = z_i - \bar{z}$. The leave-one-out mean $\bar{z}_{-i} = \frac{1}{n}\sum_{j \neq i} z_i$ is equal to $\bar{z} - \frac{z_i}{n}$. The IJ estimate $\tilde{\beta}^{\mathrm{IJ}}_{\text{-}i}(\lambda)$ is given by

$$\tilde{\beta}^{\mathrm{IJ}}_{\text{-}i}(\lambda) = \begin{cases} 0 & \lambda \geq \bar{z} \\ \hat{\beta}(\lambda) - \frac{z_i - \hat{\beta}(\lambda)}{n} = \bar{z} - \lambda - \frac{\epsilon_i + \lambda}{n} & \text{else.} \end{cases}$$

The IJ approximate cross-validation estimate for this estimator is therefore given by

$$2\mathbf{ACV}^{\mathrm{IJ}}(\lambda) = \mathbb{E}_n[(z_i - \tilde{\beta}^{\mathrm{IJ}}_{\text{-}i}(\lambda))^2] = \begin{cases} \bar{z}^2 + 1 & \lambda \geq \bar{z} \\ (\lambda^2 + 1)(1 + \frac{1}{n})^2 & \text{else.} \end{cases}$$

By construction of our dataset, $\lambda \geq 0 > -\bar{z}_{-i}$ for all i. Now, the leave-one-out estimator of β is given by $\hat{\beta}_{-i}(\lambda) = \max\Big(\bar{z} - \frac{n}{n-1}\lambda - \frac{1}{n-1}\epsilon_i, 0\Big)$, and the leave-one-out CV estimate is given by

$$
\begin{aligned}
&2\mathbf{CV}(\lambda) = \mathbb{E}_n[(\bar{z} - z_i)^2] + (\bar{z} - \hat{\beta}(\lambda))^2 \\
&+ 2\mathbb{E}_n[(\hat{\beta}_{-i}(\lambda) - \hat{\beta}(\lambda))z_i] + \mathbb{E}_n[(\hat{\beta}_{-i}(\lambda) - \hat{\beta}(\lambda))^2] \\
&= 1 + \min(\bar{z}, \lambda)^2 \\
&+ \mathbb{E}_n[\max(\Big(\bar{z} - \tfrac{n}{n-1}\lambda - \tfrac{1}{n-1}\epsilon_i, 0\Big) - (\max(\bar{z} - \lambda, 0)) \cdot (\bar{z} + \epsilon_i)] \\
&+ \mathbb{E}_n[(\max\Big(\bar{z} - \tfrac{n}{n-1}\lambda - \tfrac{1}{n-1}\epsilon_i, 0\Big) - (\max(\bar{z} - \lambda, 0))^2].
\end{aligned}
$$

Evaluating these expressions at $\lambda = \bar{z}$, we get $\hat{\beta}_{-i}(\bar{z}) = \max\Big(-\frac{z_i}{n-1}, 0\Big)$, so that

$$
\begin{aligned}
2\mathbf{ACV}^{\mathrm{IJ}}(\bar{z}) &= \bar{z}^2 + 1, \\
2\mathbf{CV}(\bar{z}) &= 1 + \bar{z}^2 + \mathbb{E}_n[\max(-\tfrac{z_i}{n-1}, 0) \cdot z_i] \\
&\quad + \mathbb{E}_n[\max(-\tfrac{z_i}{n-1}, 0)^2]
\end{aligned}
$$

and thus

$$
\begin{aligned}
&2(\mathbf{ACV}^{\mathrm{IJ}}(\bar{z}) - \mathbf{CV}(\bar{z})) \\
&= \mathbb{E}_n[\max(-\tfrac{z_i}{n-1}, 0) \cdot z_i] + \mathbb{E}_n[\max(-\tfrac{z_i}{n-1}, 0)^2] \\
&= \mathbb{E}_n[z_i^2 \cdot \mathbf{1}(z_i < 0)] \cdot \frac{n}{(n-1)^2}.
\end{aligned}
$$

Our dataset was constructed such that

$$
\begin{aligned}
\mathbb{P}_n(\epsilon_i < 0) &= 1/2 \\
\mathbb{E}_n[\epsilon_i | \epsilon_i < 0] &= \sqrt{2/\pi} \\
\mathbb{E}_n[\epsilon_i^2 | \epsilon_i < 0] &= 1 \\
\mathbb{E}_n[z_i^2 \cdot \mathbf{1}(z_i < 0)] &= \mathbb{E}_n[(\bar{z}^2 + 2\bar{z}\epsilon_i + \epsilon_i^2) \cdot \mathbf{1}(\epsilon_i < 0)] \\
&= \tfrac{1}{2}(\bar{z}^2 - 2\bar{z}\sqrt{2/\pi} + 1),
\end{aligned}
$$

and thus

$$
\mathbf{ACV}^{\mathrm{IJ}}(\bar{z}) - \mathbf{CV}(\bar{z}) = \tfrac{n}{4(n-1)^2}\Big(1 - 2\bar{z}\sqrt{2/\pi} + \bar{z}^2\Big).
$$

To make $\lambda = \bar{z}$ the $\mathbf{ACV}^{\mathrm{IJ}}$ minimizing choice in this example (a condition we have not assumed thus far), it suffices to have $\bar{z} \leq \sqrt{2/n}$. For the choice $\bar{z} = \sqrt{2/n}$, we get

$$
\mathbf{ACV}^{\mathrm{IJ}}(\bar{z}) - \mathbf{CV}(\bar{z}) = \frac{n}{4(n-1)^2}\Big(1 - \frac{4}{\sqrt{n\pi}} + \tfrac{2}{n}\Big).
$$

H Proof of Prop. 9

We write $\mathbb{E}_n[z_i] \triangleq \frac{1}{n}\sum_{i=1}^n z_i$ to denote a sample average. Consider the patched Lasso estimator

$$
\hat{\beta}(\lambda) \triangleq \operatorname*{argmin}_{\beta} \tfrac{1}{2n}\sum_i (\beta - z_i)^2 + \lambda \min(|\beta|, \tfrac{\delta}{2} + \tfrac{\beta^2}{2\delta}) = \max\Big(\bar{z} - \lambda, \tfrac{\bar{z}}{1+\lambda/\delta}\Big).
$$

Define $\epsilon_i = z_i - \bar{z}$. The leave-one-out mean is equal to $\bar{z}_{-i} = \bar{z} - \frac{\epsilon_i}{n}$. For $\epsilon_i/n < \bar{z}$, we have $\hat{\beta}_{-i}(\lambda) - \hat{\beta}(\lambda) = -\frac{\epsilon_i}{n-1}$ if $\epsilon_i < 0$ and $\hat{\beta}_{-i}(\lambda) - \hat{\beta}(\lambda) = -\frac{\epsilon_i}{n-1}\frac{1}{1+\lambda/\delta}$ if $\epsilon_i > 0$. Considering left hand derivatives, we also have $\tilde{\beta}^{\mathrm{IJ}}_{-i}(\lambda) - \hat{\beta}(\lambda) = -\frac{\epsilon_i}{n}$.

For $\lambda = \delta$ and $\bar{z} = 2\delta$, we get $\hat{\beta}(\lambda) = \delta$, and and thus, by our choice of dataset,

$$
\begin{aligned}
\mathbf{ACV}^{\mathrm{IJ}}(\bar{z}) - \mathbf{CV}(\bar{z}) &= \tfrac{1}{2}\mathbb{E}_n\Big[(Z_i - \tilde{\beta}^{\mathrm{IJ}}_{\text{-}i}(\lambda))^2 - (Z_i - \hat{\beta}_{\text{-}i}(\lambda))^2\Big] \\
&= \tfrac{1}{2}\mathbb{E}_n\Big[(2\delta + \epsilon_i - \delta + \tfrac{\epsilon_i}{n})^2 - (2\delta + \epsilon_i - \delta + \tfrac{1}{2}\tfrac{\epsilon_i}{n-1})^2 | \epsilon_i > 0\Big] \\
&\quad + \tfrac{1}{2}\mathbb{E}_n\Big[(2\delta + \epsilon_i - \delta + \tfrac{\epsilon_i}{n})^2 - (2\delta + \epsilon_i - \delta + \tfrac{\epsilon_i}{n-1})^2 | \epsilon_i < 0\Big] \\
&= \tfrac{1}{2}\mathbb{E}_n\Big[(\delta + \epsilon_i(1 + \tfrac{1}{n}))^2 - (\delta + \epsilon_i(1 + \tfrac{1}{2(n-1)}))^2 | \epsilon_i > 0\Big] + O(\tfrac{1}{n^2}) \\
&= \delta \mathbb{E}_n[\epsilon_i | \epsilon_i > 0] \cdot \tfrac{1}{n} + O(\tfrac{1}{n^2}) \\
&= \delta\sqrt{2/\pi} \cdot \tfrac{1}{n} + O(\tfrac{1}{n^2}).
\end{aligned}
$$

I Proof of Thm. 10: ProxACV-CV assessment error

The optimization perspective adopted in this paper naturally points towards generalizations of the proximal estimator (13). In particular, stronger assessment guarantees can be provided for (regularized) higher-order Taylor approximations of the objective function. For example, for $p \geq 2$, we may define the p-th order approximation

$$
\begin{aligned}
&\mathbf{ProxACV}^{\mathrm{HO}}_p(\lambda) \triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(z_i, \tilde{\beta}^{\text{Prox-HO}_p}_{\text{-}i}(\lambda)) \quad \text{with} \\
&\quad \tilde{\beta}^{\text{Prox-HO}_p}_{\text{-}i}(\lambda) \triangleq \operatorname{argmin}_{\beta \in \mathbb{R}^d}\{\hat{\ell}_p(\mathbb{P}_{n,\text{-}i}, \beta; \hat{\beta}(\lambda)) + \lambda\pi(\beta)\}
\end{aligned}
$$

which recovers **ProxACV** (12) and the **ProxACV** estimator (13) in the setting $p = 2$. We also define the regularized p-th order approximation

$$
\begin{aligned}
&\mathbf{ProxACV}^{\mathrm{RHO}}_p(\lambda) \triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(z_i, \tilde{\beta}^{\text{Prox-RHO}_p}_{\text{-}i}(\lambda)) \quad \text{with} \\
&\quad \tilde{\beta}^{\text{Prox-RHO}_p}_{\text{-}i}(\lambda) \triangleq \operatorname{argmin}_{\beta \in \mathbb{R}^d}\Big\{\hat{\ell}_p(\mathbb{P}_{n,\text{-}i}, \beta; \hat{\beta}(\lambda)) + \tfrac{\mathrm{Lip}(\nabla^p_\beta \ell(\mathbb{P}_{n,\text{-}i}, \cdot))}{p+1}\|\hat{\beta}(\lambda) - \beta\|_2^{p+1} + \lambda\pi(\beta)\Big\},
\end{aligned}
$$

where $\hat{\ell}_p(\mathbb{P}_{n,\text{-}i}, \cdot; \hat{\beta}(\lambda))$ is a p-th order Taylor expansion of the loss $\ell_p(\mathbb{P}_{n,\text{-}i}, \cdot)$ about $\hat{\beta}(\lambda)$, that is, $\hat{f}_p(\beta; \hat{\beta}(\lambda)) \triangleq \sum_{k=0}^p \frac{1}{k!}\nabla^k f(\hat{\beta}(\lambda))(\beta - \hat{\beta}(\lambda))^{\otimes k}$. To analyze both, we will make use of the following assumptions which generalize Assump. 1c.

Assumption 1g (Curvature of proximal Taylor approximation)**.** *For some $p, q, c_m > 0$, all $i \in [n]$, and all λ in a given $\Lambda \subseteq [0, \infty]$, $\widehat{\ell}_p(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda)) + \lambda\pi$ has $\nu(r) = c_m r^q$ gradient growth.*

Assumption 1h (Curvature of regularized proximal Taylor approximation)**.** *For some $p, q, c_m > 0$, all $i \in [n]$, and all λ in a given $\Lambda \subseteq [0, \infty]$, $\widehat{\ell}_p(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda)) + \frac{\mathrm{Lip}(\nabla^p_\beta \ell(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda))}{p+1}\|\cdot - \hat{\beta}(\lambda)\|_2^{p+1} + \lambda\pi$ has $\nu(r) = c_m r^q$ gradient growth.*

Thm. 10 will then follow from the following more general statement, proved in App. I.1.

Theorem 29 (**ProxACV**$^{\mathrm{HO}}_p$**-CV** and **ProxACV**$^{\mathrm{RHO}}_p$**-CV** assessment error)**.** *If Assumps. 3c, 1d, and 1g hold for some $\Lambda \subseteq [0, \infty]$, then, for all $\lambda \in \Lambda$ and $i \in [n]$,*

$$
\|\tilde{\beta}^{\textit{Prox-HO}_p}_{\text{-}i}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2^{q-1} \leq \tilde{\kappa}_p\|\hat{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2^p \quad \textit{with} \quad \tilde{\kappa}_p \triangleq \tfrac{C_{\ell,p+1}}{p!\,c_m}. \tag{32a}
$$

If Assumps. 3c, 1d, and 1g hold for some $\Lambda \subseteq [0, \infty]$, then, for all $\lambda \in \Lambda$ and $i \in [n]$,

$$
\|\tilde{\beta}^{\textit{Prox-RHO}_p}_{\text{-}i}(\lambda) - \hat{\beta}_{\text{-}i}(\lambda)\|_2^{q-1} \leq 2\tilde{\kappa}_p\|\hat{\beta}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2^p. \tag{32b}
$$

If Assumps. 2, 3c, 1d, and 1g hold for some $\Lambda \subseteq [0, \infty]$ and each $(s, r) \in \{(0, \frac{p+(q-1)^2}{(q-1)^2}), (1, \frac{2p}{(q-1)^2}), (1, \frac{p+q-1}{(q-1)^2})\}$, then, for all $\lambda \in \Lambda$,

$$
\begin{aligned}
&|\mathbf{ProxACV}^{HO}_p(\lambda) - \mathbf{CV}(\lambda)| \\
&\leq \frac{1}{n^{\frac{p}{(q-1)^2}}}\frac{(\tilde{\kappa}_p)^{\frac{1}{q-1}}}{c_m^{\frac{p}{(q-1)^2}}}\mathrm{B}^\ell_{0,\frac{p+(q-1)^2}{(q-1)^2}} + \frac{1}{2}\frac{1}{n^{\frac{2p}{(q-1)^2}}}\frac{(\tilde{\kappa}_p)^{\frac{2}{q-1}}}{c_m^{\frac{2p}{(q-1)^2}}}\mathrm{B}^\ell_{1,\frac{2p}{(q-1)^2}} + \frac{1}{n^{\frac{p+q-1}{(q-1)^2}}}\frac{(\tilde{\kappa}_p)^{\frac{1}{q-1}}}{c_m^{\frac{p+q-1}{(q-1)^2}}}\mathrm{B}^\ell_{1,\frac{p+q-1}{(q-1)^2}} \quad \textit{and}
\end{aligned} \tag{33a}
$$

If Assumps. 2, 3c, 1d, and 1h holds for Λ *and each* $(s,r) \in \{(0, \frac{p+(q-1)^2}{(q-1)^2}), (1, \frac{2p}{(q-1)^2}), (1, \frac{p+q-1}{(q-1)^2})\}$*, then, for all* $\lambda \in \Lambda$*,*

$$\begin{aligned}&|\mathbf{ProxACV}_p^{RHO}(\lambda) - \mathbf{CV}(\lambda)| \\ &\leq \frac{1}{n^{\frac{p}{(q-1)^2}}} \frac{(2\tilde{\kappa}_p)^{\frac{1}{q-1}}}{c_m^{\frac{p}{(q-1)^2}}} \mathrm{B}^{\ell}_{0,\frac{p+(q-1)^2}{(q-1)^2}} + \frac{1}{2} \frac{1}{n^{\frac{2p}{(q-1)^2}}} \frac{(2\tilde{\kappa}_p)^{\frac{2}{q-1}}}{c_m^{\frac{2p}{(q-1)^2}}} \mathrm{B}^{\ell}_{1,\frac{2p}{(q-1)^2}} + \frac{1}{n^{\frac{p+q-1}{(q-1)^2}}} \frac{(2\tilde{\kappa}_p)^{\frac{1}{q-1}}}{c_m^{\frac{p+q-1}{(q-1)^2}}} \mathrm{B}^{\ell}_{1,\frac{p+q-1}{(q-1)^2}}.\end{aligned} \tag{33b}$$

Thm. 10 follows from Thm. 29 with $p = q = 2$ since Assump. 1c (with $0 \in \Lambda$) implies $\mu = c_m$ strong convexity for $\widehat{\ell}_2(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda))$. Since π is convex, we further have μ strong convexity and hence $\nu(r) = \mu r^2$ gradient growth for $\widehat{\ell}_2(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda)) + \lambda\pi(\cdot)$ for each $\lambda \in \Lambda$.

I.1 Proof of Thm. 29: $\mathbf{ProxACV}_p^{\mathbf{HO}}$-CV and $\mathbf{ProxACV}_p^{\mathbf{RHO}}$-CV assessment error

I.1.1 Proof of (32a) and (32b): Proximity of $\mathbf{ProxACV}_p^{\mathbf{HO}}$, $\mathbf{ProxACV}_p^{\mathbf{RHO}}$, and CV estimators

The proofs follow exactly as in Apps. B.1.2 and B.1.3 if we take $\varphi(x) = \ell(x)$, $\widehat{\varphi}_p(x; w) = \widehat{\ell}_p(x; w)$, $\varphi_0(x) = \pi(x)$, and $w = \hat{\beta}(\lambda)$ and invoke Assumps. 3c, 1g, and 1h in place of Assumps. 3b, 1e, and 1f, respectively.

I.1.2 Proof of (33a) and (33b): Proximity of $\mathbf{ProxACV}_p^{\mathbf{HO}}$, $\mathbf{ProxACV}_p^{\mathbf{RHO}}$, and CV

This proofs follow exactly as in follows directly from the proof contained in Apps. B.1.4 and B.1.5 if we substitute (32a) and (32b) for (16a) and (16b) respectively.

J Proof of Thm. 11: $\mathbf{ProxACV}^{\mathbf{IJ}}$-ProxACV assessment error

We will prove the following more detailed statement from which Thm. 11 immediately follows.

Theorem 30 ($\mathbf{ProxACV}^{\mathrm{IJ}}$-ProxACV assessment error**).** *If Assump. 1c holds for* $\Lambda \subseteq [0, \infty]$ *with* $0 \in \Lambda$*, then, for each* $\lambda \in \Lambda$*,*

$$\|\tilde{\beta}^{prox,IJ}_{\text{-}i}(\lambda) - \tilde{\beta}^{prox}_{\text{-}i}(\lambda)\|_2 \leq \frac{\|\nabla^2_\beta \ell(z_i, \hat{\beta}(\lambda))\|_{\mathrm{op}} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2}{c_m^2 n^2} \tag{34}$$

If, in addition, Assump. 2 holds for Λ *and each* $(s,r) \in \{(1,2),(2,2),(3,2)\}$*, then*

$$|\mathbf{ProxACV}^{\mathrm{IJ}}(\lambda) - \mathbf{ProxACV}(\lambda)| \leq \frac{1}{n^2 c_m^2} \mathrm{B}^{\ell}_{1,2} + \frac{1}{2n^4 c_m^4} \mathrm{B}^{\ell}_{3,2} + \frac{1}{n^3 c_m^3} \mathrm{B}^{\ell}_{2,2}. \tag{35}$$

J.1 Proof of (34): Proximity of $\mathbf{ProxACV}^{\mathbf{IJ}}$ and ProxACV estimators

The concavity of the minimum eigenvalue, Jensen's inequality, and Assump. 1c with $0 \in \Lambda$ imply that

$$\mathrm{mineig}(\mathrm{H}_\ell) = \mathrm{mineig}(\tfrac{n}{n-1} \tfrac{1}{n} \textstyle\sum_{i=1}^n \mathrm{H}_{\ell,i}) \geq \tfrac{n}{n-1} \tfrac{1}{n} \textstyle\sum_{i=1}^n \mathrm{mineig}(\mathrm{H}_{\ell,i}) \geq \tfrac{n}{n-1} c_m$$

for $\mathrm{H}_\ell = \nabla^2_\beta \ell(\mathbb{P}_n, \hat{\beta}(\lambda))$ and $\mathrm{H}_{\ell,i} = \nabla^2_\beta \ell(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda))$. Moreover, Assump. 1c with $0 \in \Lambda$ implies $\mu = c_m$ strong convexity for $\widehat{\ell}_2(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda))$; since π is convex, we further have μ strong convexity and hence $\nu(r) = \mu r^2$ gradient growth for $\widehat{\ell}_2(\mathbb{P}_{n,\text{-}i}, \cdot, \lambda; \hat{\beta}(\lambda)) + \lambda\pi(\cdot)$ for each $\lambda \in \Lambda$. Hence, we may apply the Proximal Newton Comparison Lemma 20 to obtain

$$\begin{aligned}\|\tilde{\beta}^{\mathrm{prox,IJ}}_{\text{-}i}(\lambda) - \tilde{\beta}^{\mathrm{prox}}_{\text{-}i}(\lambda)\|_2 &\leq \tfrac{1}{c_m} \|\nabla^2_\beta \ell(\mathbb{P}_n, \hat{\beta}(\lambda)) - \nabla^2_\beta \ell(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda))\|_{\mathrm{op}} \|\tilde{\beta}^{\mathrm{prox}}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2 \\ &\leq \tfrac{1}{n} \tfrac{1}{c_m} \|\nabla^2_\beta \ell(z_i, \hat{\beta}(\lambda))\|_{\mathrm{op}} \|\tilde{\beta}^{\mathrm{prox}}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2.\end{aligned}$$

To complete the bound, we note that $\hat{\beta}(\lambda) = \mathrm{prox}^{\mathrm{H}_{\ell,i}}_{\lambda\pi}(\hat{\beta}(\lambda) - \mathrm{H}_{\ell,i}^{-1} \nabla_\beta \ell(\mathbb{P}_n, \hat{\beta}(\lambda)))$ and use the 1-Lipschitzness of the prox operator to conclude that

$$\|\tilde{\beta}^{\mathrm{prox}}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2 \leq \|\mathrm{H}_{\ell,i}^{-1}(\nabla_\beta \ell(\mathbb{P}_n, \hat{\beta}(\lambda)) - \nabla_\beta \ell(\mathbb{P}_{n,\text{-}i}, \hat{\beta}(\lambda)))\|_2 \leq \tfrac{1}{nc_m} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2. \tag{36}$$

J.2 Proof of (35): Proximity of ProxACV$^{\textbf{IJ}}$ and ProxACV

Fix any $\lambda \in \Lambda$. To control the discrepancy between $\mathbf{ProxACV}(\lambda)$ and $\mathbf{ProxACV}^{\text{IJ}}(\lambda)$, we first rewrite the difference using Taylor's theorem with Lagrange remainder:

$$\begin{aligned}\mathbf{ProxACV}(\lambda) - \mathbf{ProxACV}^{\text{IJ}}(\lambda) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(z_i, \tilde\beta^{\text{prox}}_{\text{-}i}(\lambda)) - \ell(z_i, \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda))\\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat\beta_{\text{-}i}(\lambda)), \tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda)\rangle\\ &\quad + \tfrac{1}{2}\nabla^2_\beta \ell(z_i, \tilde s_i)[\tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda)]^{\otimes 2}\end{aligned}$$

for some $\tilde s_i \in \{t\tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) + (1-t)\tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda) : t \in [0,1]\}$. We next use the mean-value theorem to expand each function $\langle \nabla_\beta \ell(z_i, \cdot), \tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda)\rangle$ around the full-data estimator $\hat\beta(\lambda)$:

$$\begin{aligned}\mathbf{ProxACV}(\lambda) - \mathbf{ProxACV}^{\text{IJ}}(\lambda) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat\beta(\lambda)), \tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda)\rangle\\ &\quad + \tfrac{1}{2}\nabla^2_\beta \ell(z_i, \tilde s_i)[\tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda)]^{\otimes 2}\\ &\quad + \langle \nabla^2_\beta \ell(z_i, s_i)(\hat\beta_{\text{-}i}(\lambda) - \hat\beta(\lambda)), \tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda)\rangle\end{aligned}$$

for some $s_i \in \{t\hat\beta(\lambda) + (1-t)\tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda) : t \in [0,1]\}$. Finally, we invoke Cauchy-Schwarz, the definition of the operator norm, the estimator proximity results (15) and (34), and Assump. 2 to obtain

$$\begin{aligned}|\mathbf{ProxACV}(\lambda) - \mathbf{ProxACV}^{\text{IJ}}(\lambda)| &\le \tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla_\beta \ell(z_i, \hat\beta(\lambda))\|_2 \|\tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda)\|_2\\ &\quad + \tfrac{1}{2}\|\nabla^2_\beta \ell(z_i, \tilde s_i)\|_{\text{op}} \|\tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda)\|_2^2\\ &\quad + \|\nabla^2_\beta \ell(z_i, s_i)\|_{\text{op}} \|\hat\beta_{\text{-}i}(\lambda) - \hat\beta(\lambda)\|_2 \|\tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \tilde\beta^{\text{prox,IJ}}_{\text{-}i}(\lambda))\|_2\\ &\le \tfrac{1}{n^2 c_m^2}\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla^2_\beta \ell(z_i, \hat\beta(\lambda))\|_{\text{op}} \|\nabla_\beta \ell(z_i, \hat\beta(\lambda))\|_2^2\\ &\quad + \tfrac{1}{2n^4 c_m^4}\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla^2_\beta \ell(z_i, \tilde s_i)\|_{\text{op}} \|\nabla^2_\beta \ell(z_i, \hat\beta(\lambda))\|_{\text{op}}^2 \|\nabla_\beta \ell(z_i, \hat\beta(\lambda))\|_2^2\\ &\quad + \tfrac{1}{n^3 c_m^3}\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla^2_\beta \ell(z_i, \tilde s_i)\|_{\text{op}} \|\nabla^2_\beta \ell(z_i, \hat\beta(\lambda))\|_{\text{op}} \|\nabla_\beta \ell(z_i, \hat\beta(\lambda))\|_2^2\\ &\le \tfrac{1}{n^2 c_m^2}\mathrm{B}^\ell_{1,2} + \tfrac{1}{2n^4 c_m^4}\mathrm{B}^\ell_{3,2} + \tfrac{1}{n^3 c_m^3}\mathrm{B}^\ell_{2,2}.\end{aligned}$$

K Proof of Thm. 12: ProxACV-CV selection error

The first claim follows immediately from the following more detailed version of Thm. 12.

Theorem 31 (Weak **ProxACV** proximity implies $\hat\beta$ proximity)**.** *Suppose Assumps. 1c and 3c hold for some $\Lambda \subseteq [0,\infty]$ and each $(s,r) \in \{(0,2),(1,2)\}$. Then, for all $\lambda', \lambda \in \Lambda$ with $\lambda' < \lambda$,*

$$\|\hat\beta(\lambda) - \hat\beta(\lambda')\|_2^2 \le C_{1,\lambda,\lambda'}\big(\tfrac{4\mathrm{B}^\ell_{0,2}}{nc_m} + \tfrac{\mathrm{B}^\ell_{1,2}}{n^2 c_m^2} + \mathbf{ProxACV}(\lambda) - \mathbf{ProxACV}(\lambda')\big), \tag{37}$$

where $C_{1,\lambda,\lambda'} = \frac{2}{c_m}\frac{\lambda-\lambda'}{\lambda+\lambda'}\frac{n-1}{n}$.

Proof Fix any $\lambda', \lambda \in \Lambda$ with $\lambda' < \lambda$. We begin by writing the difference in estimator training losses as a difference in **ProxACV** values plus a series of error terms:

$$\ell(\mathbb{P}_n, \hat\beta(\lambda)) - \ell(\mathbb{P}_n, \hat\beta(\lambda')) = \mathbf{ProxACV}(\lambda) - \mathbf{ProxACV}(\lambda') + \Delta T_1 - \Delta T_2 - \Delta T_3 \tag{38}$$

for

$$\begin{aligned}\Delta T_1 &\triangleq \widehat{\mathbf{ProxACV}}(\lambda) - \mathbf{ProxACV}(\lambda) + \mathbf{ProxACV}(\lambda') - \widehat{\mathbf{ProxACV}}(\lambda')\\ \widehat{\mathbf{ProxACV}}(\lambda) &\triangleq \ell(\mathbb{P}_n, \hat\beta(\lambda)) + \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat\beta(\lambda)), \tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \hat\beta(\lambda)\rangle\\ \Delta T_2 &\triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat\beta(\lambda)) - \nabla_\beta \ell(z_i, \hat\beta(\lambda')), \tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \hat\beta(\lambda)\rangle\\ \Delta T_3 &\triangleq \tfrac{1}{n}\textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat\beta(\lambda')), \tilde\beta^{\text{prox}}_{\text{-}i}(\lambda) - \hat\beta(\lambda) - (\tilde\beta^{\text{prox}}_{\text{-}i}(\lambda') - \hat\beta(\lambda'))\rangle.\end{aligned}$$

We will frequently use the bound (36) which follows from Assump. 1c and implies

$$\|\hat\beta(\lambda) - \tilde\beta^{\text{prox}}_{\text{-}i}(\lambda)\|_2 \le \tfrac{1}{n}\tfrac{1}{c_m}\|\nabla_\beta \ell(z_i, \hat\beta(\lambda))\|_2$$

for each $i \in [n]$.

To bound ΔT_1, we first employ Taylor's Theorem with Lagrange remainder,

$$\mathbf{ProxACV}(\lambda) - \widehat{\mathbf{ProxACV}}(\lambda) = \tfrac{1}{2n} \textstyle\sum_{i=1}^n \nabla_\beta^2 \ell(z_i, s_i)[\tilde{\beta}^{\text{prox}}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)]^{\otimes 2}$$

for some $s_i \in \mathcal{L}_i = \{t\hat{\beta}(\lambda) + (1-t)\hat{\beta}_{\text{-}i}(\lambda) : t \in [0,1]\}$. Next we apply the definition of the operator norm, the bound (36), and Assump. 2 with $(s,r) = (1,2)$

$$\begin{aligned} |\mathbf{ProxACV}(\lambda) - \widehat{\mathbf{ProxACV}}(\lambda)| &\leq \tfrac{1}{2n} \textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\text{op}} \|\tilde{\beta}^{\text{prox}}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2^2 \\ &\leq \tfrac{1}{2}\tfrac{1}{n} \textstyle\sum_{i=1}^n \|\nabla_\beta^2 \ell(z_i, s_i)\|_{\text{op}} \Big(\tfrac{1}{n^2}\tfrac{1}{c_m^2} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^2\Big) \\ &\leq \tfrac{1}{n^2}\tfrac{1}{2c_m^2} \mathrm{B}^\ell_{1,2} \end{aligned}$$

Since an identical bound holds for λ', we have

$$|\Delta T_1| \leq \tfrac{1}{n^2}\tfrac{1}{c_m^2} \mathrm{B}^\ell_{1,2}.$$

To bound ΔT_2 and ΔT_3, we apply Cauchy-Schwarz, the triangle inequality, the bound (36), the arithmetic-geometric mean inequality, and Assump. 2 with $(s,r) = (0,2)$ to find

$$\begin{aligned} |\Delta T_2| &= \tfrac{1}{n}|\tfrac{1}{n} \textstyle\sum_{i=1}^n \langle \nabla_\beta \ell(z_i, \hat{\beta}(\lambda)) - \nabla_\beta \ell(z_i, \hat{\beta}(\lambda')), \tilde{\beta}^{\text{prox}}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\rangle| \\ &\leq \tfrac{1}{n} \textstyle\sum_{i=1}^n \|\tilde{\beta}^{\text{prox}}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2 \Big(\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 + \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2\Big) \\ &\leq \tfrac{1}{n} \textstyle\sum_{i=1}^n \tfrac{1}{n}\tfrac{1}{c_m} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 \Big(\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 + \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2\Big) \\ &\leq \tfrac{1}{n}\tfrac{1}{c_m}\tfrac{1}{n} \textstyle\sum_{i=1}^n \tfrac{3}{2}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2^2 + \tfrac{1}{2}\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2^2 \leq \tfrac{2}{n}\tfrac{\mathrm{B}^\ell_{0,2}}{c_m} \quad \text{and} \\ |\Delta T_3| &\leq \tfrac{1}{n} \textstyle\sum_{i=1}^n \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2 \Big(\|\tilde{\beta}^{\text{prox}}_{\text{-}i}(\lambda) - \hat{\beta}(\lambda)\|_2 + \|\tilde{\beta}^{\text{prox}}_{\text{-}i}(\lambda') - \hat{\beta}(\lambda')\|_2\Big) \\ &\leq \tfrac{1}{n} \textstyle\sum_{i=1}^n \tfrac{1}{n}\tfrac{1}{c_m} \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 \Big(\|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))\|_2 + \|\nabla_\beta \ell(z_i, \hat{\beta}(\lambda'))\|_2\Big) \leq \tfrac{2}{n}\tfrac{\mathrm{B}^\ell_{0,2}}{c_m}. \end{aligned}$$

The advertised result (37) now follows by combining Lemma 27 with the loss difference decomposition (38) and the component $\Delta T_1, \Delta T_2$, and ΔT_3 bounds. □

The second claim in Thm. 12 follows from Thm. 31 and the following lemma.

Lemma 32. *If Assumps. 1c, 2, and 3c hold for some $\Lambda \subseteq [0,\infty]$ and each $(s,r) \in \{(0,3),(1,3),(1,4)\}$. If $\lambda_{\mathbf{ProxACV}} \in \operatorname{argmin}_{\lambda\in\Lambda} \mathbf{ProxACV}(\lambda)$ and $\lambda_{\mathbf{CV}} \in \operatorname{argmin}_{\lambda\in\Lambda} \mathbf{CV}(\lambda)$, then*

$$0 \leq \mathbf{ProxACV}(\lambda_{\mathbf{CV}}) - \mathbf{ProxACV}(\lambda_{\mathbf{ACV}}) \leq \tfrac{C_{\ell,3}}{n^2}\big(\tfrac{\mathrm{B}^\ell_{0,3}}{c_m^3} + \tfrac{\mathrm{B}^\ell_{1,3}}{nc_m^4} + \tfrac{C_{\ell,3}\mathrm{B}^\ell_{1,4}}{4n^2c_m^6}\big).$$

Proof Since $\lambda_{\mathbf{CV}}$ minimizes $\mathbf{CV}$ and $\lambda_{\mathbf{ProxACV}}$ minimizes $\mathbf{ProxACV}$,

$$\begin{aligned} 0 &\leq \mathbf{ProxACV}(\lambda_{\mathbf{CV}}) - \mathbf{ProxACV}(\lambda_{\mathbf{ProxACV}}) \\ &\leq \mathbf{ProxACV}(\lambda_{\mathbf{CV}}) - \mathbf{ProxACV}(\lambda_{\mathbf{ProxACV}}) + \mathbf{CV}(\lambda_{\mathbf{ProxACV}}) - \mathbf{CV}(\lambda_{\mathbf{CV}}). \end{aligned}$$

The result now follows from two applications of Thm. 10. □

L Proof of Prop. 13: $O(1/\sqrt{n})$ error bound is tight

For each $i \in [n]$, define $\bar{z}_i = \bar{z} - \frac{1}{n}z_i$. For the target objective,

$$\mathbf{ProxACV}(\lambda) = \mathbf{CV}(\lambda) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{1}{2}(\hat{\beta}_{\text{-}i}(\lambda) - z_i)^2 \quad \text{for all} \quad \lambda \in [0,\infty],$$

and a straightforward calculation shows

$$\hat{\beta}(\lambda) = \begin{cases} \bar{z} - \lambda & \text{if } \bar{z} > \lambda \\ \bar{z} + \lambda & \text{if } \bar{z} < -\lambda \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \hat{\beta}_{\text{-}i}(\lambda) = \begin{cases} \bar{z}_i - \lambda & \text{if } \bar{z}_i > \lambda \\ \bar{z}_i + \lambda & \text{if } \bar{z}_i < -\lambda \\ 0 & \text{otherwise} \end{cases}$$

for each $i \in [n]$. Hence, for each i, $\hat{\beta}_{\text{-}i}(0) = \bar{z}_i$,

$$\hat{\beta}_{\text{-}i}(\bar{z}) = \begin{cases} -\frac{1}{n} z_i & \text{if } 0 > z_i \\ \bar{z}_i + \bar{z} & \text{if } 2n\bar{z} < z_i \,, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \hat{\beta}_{\text{-}i}(\bar{z}) - z_i = \begin{cases} -\frac{n+1}{n} z_i & \text{if } 0 > z_i \\ \bar{z}_i - z_i + \bar{z} & \text{if } 2n\bar{z} < z_i \,. \\ -z_i & \text{otherwise} \end{cases}$$

Let $C_1 = \{i \in [n] \,:\, z_i \le 0\}$, $C_2 = \{i \in [n] \,:\, z_i > 2\sqrt{2n} = 2n\bar{z}\}$ and $C_3 = \{i \in [n] \,:\, z_i \notin C_1 \cup C_2\}$. We have selected our dataset so that C_2 is empty. Therefore

$$2\textbf{ProxACV}(\bar{z}) = \tfrac{1}{n} \textstyle\sum_{i \in C_1} \tfrac{(n+1)^2}{n^2} z_i^2 + \tfrac{1}{n} \sum_{i \in C_3} z_i^2 = \tfrac{(n+1)^2}{n^2} \tfrac{1}{n} \sum_{i=1}^n z_i^2 - (\tfrac{2}{n} + \tfrac{1}{n^2}) \tfrac{1}{n} \sum_{i \in C_3} z_i^2.$$

Meanwhile,

$$\begin{aligned} 2\textbf{ProxACV}(0) &= \tfrac{1}{n} \textstyle\sum_{i=1}^n (\bar{z}_i - z_i)^2 \\ &= \tfrac{1}{n} \textstyle\sum_{i=1}^n (\bar{z} - (1 + 1/n)\bar{z})^2 + (1 + 1/n)^2 (\bar{z} - z_i)^2 \\ &= \tfrac{1}{n^2} \bar{z}^2 + \tfrac{(n+1)^2}{n^2} \tfrac{1}{n} \textstyle\sum_{i=1}^n z_i^2 - \tfrac{(n+1)^2}{n^2} \bar{z}^2 \\ &= \tfrac{(n+1)^2}{n^2} \tfrac{1}{n} \textstyle\sum_{i=1}^n z_i^2 - (1 + \tfrac{2}{n}) \bar{z}^2. \end{aligned}$$

Hence,

$$2\textbf{ProxACV}(0) - 2\textbf{ProxACV}(\bar{z}) = (\tfrac{2}{n} + \tfrac{1}{n^2}) \tfrac{1}{n} \textstyle\sum_{i \in C_3} z_i^2 - (1 + \tfrac{2}{n}) \bar{z}^2 = (\tfrac{2}{n} + \tfrac{1}{n^2}) \tfrac{a^2}{2} - (1 + \tfrac{2}{n}) \bar{z}^2 = \tfrac{5}{n^2}.$$

M Additional Experiment Details

M.1 ProxACV versus ACV and ACV$^{\mathbf{IJ}}$

This section provides additional experimental details for the experiment of Sec. 5.1. In this experiment, we use the exact experimental setup and code of [Stephenson and Broderick, 2019, App. F] with a modified number of datapoints ($n = 150$). Specifically, we employ an ℓ_1 regularized logistic regression objective with 150 feature coefficients plus an intercept coefficient. The data matrix of covariates is generated with i.i.d. $N(0, 1)$ entries, and binary labels for each datapoint are generated independently from a logistic regression model with ground truth β^* having its first five entries drawn i.i.d. $N(0, 1)$ and the rest set to zero. We solve the proximal Newton steps for **ProxACV** using FISTA Beck and Teboulle [2009].

We compare with the non-smooth **ACV** and **ACV**$^{\mathbf{IJ}}$ extensions studied by [Obuchi and Kabashima, 2016, 2018, Rad and Maleki, 2019, Stephenson and Broderick, 2019, Wang et al., 2018] and defined by restricting $\tilde{\beta}_{\text{-}i}(\lambda)$ and $\tilde{\beta}^{\text{IJ}}_{\text{-}i}(\lambda)$ to have support only on $\hat{S} = \text{support}(\hat{\beta}(\lambda))$ and setting

$$\begin{aligned} [\tilde{\beta}_{\text{-}i}(\lambda)]^{\hat{S}} &= [\hat{\beta}(\lambda)]^{\hat{S}} + \tfrac{1}{n} [\mathrm{H}_{\ell,i}^{\hat{S},\hat{S}}]^{-1} [\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))]^{\hat{S}} \\ [\tilde{\beta}^{\text{IJ}}_{\text{-}i}(\lambda)]^{\hat{S}} &= [\hat{\beta}(\lambda)]^{\hat{S}} + \tfrac{1}{n} [\mathrm{H}_{\ell}^{\hat{S},\hat{S}}]^{-1} [\nabla_\beta \ell(z_i, \hat{\beta}(\lambda))]^{\hat{S}}, \end{aligned}$$

where $X^{\cdot,\hat{S}}$ denotes the submatrix of X with column indices in $\hat{S}$, where H_ℓ and $\mathrm{H}_{\ell,i}$ are given by (14) and (13), respectively.

M.2 ProxACV Speed-up

This section provides additional experimental details for the experiment of Sec. 5.2. For this experiment, we employed the standard graphical Lasso objective,

$$\begin{aligned} m(\mathbb{P}_n, \beta, \lambda) &= -\log\det\beta + \mathrm{tr}(\beta S) + \lambda \textstyle\sum_{j,k=1}^p |\beta_{jk}|, \\ m(\mathbb{P}_{n,\text{-}i}, \beta, \lambda) &= -\log\det\beta + \mathrm{tr}(\beta S_{-i}) + \lambda \textstyle\sum_{j,k=1}^p |\beta_{jk}|, \end{aligned}$$

where β is now a positive-definite matrix in $\mathbb{R}^{p\times p}$, $S = \frac{1}{n-1}\sum_{i=1}^{n}(z_i-\mu)(z_i-\mu)^\top$, $S_{-i} = \frac{1}{n-1}\sum_{j\neq i}(z_j-\mu_{-i})(z_j-\mu_{-i})^\top$, for $\mu = \frac{1}{n}\sum_{i=1}^{n} z_i$, and $\mu_{-i} = \frac{1}{n-1}\sum_{j\neq i} z_j$.