A Proofs for Section §3

A.1 Proof of Lemma 3.2

Goal. For any fixed $(q, C_t(q), d^+)$ and any $d^- \in C_t(q)$, show

$$\mathbf{1}\big[s_\theta(q,d^-) \ge s_\theta(q,d^+)\big] \;\le\; \frac{1}{\log 2}\,\mathcal{L}^{(1)}_{\mathrm{NCE}}\big(\theta;\,q,\,C_t(q),\,d^+\big)\,. \tag{1}$$

where

$$\begin{aligned}&\mathcal{L}^{(1)}_{\mathrm{NCE}}\big(\theta;\,q,\,C_t(q),\,d^+\big)\\ &\quad= \log\!\Big(1 + \sum_{d \in C_t(q)} \exp\!\Big(\tfrac{s_\theta(q,d) - s_\theta(q,d^+)}{\tau}\Big)\Big), \qquad \tau > 0.\end{aligned} \tag{2}$$

Averaging (1) over $(q, d^+, C_t(q), d^-)$ then yields the bound on X_t.

Step 1: Log-sum-exp dominates any single negative. Fix q, $C_t(q)$ and $d^+ \in D_q^+$. For any $d^- \in C_t(q)$, define $z \coloneqq (s_\theta(q,d^-) - s_\theta(q,d^+))/\tau$. Since all terms in the sum are nonnegative,

$$\sum_{d \in C_t(q)} e^{(s_\theta(q,d) - s_\theta(q,d^+))/\tau} \ge e^z,$$

hence

$$\mathcal{L}^{(1)}_{\mathrm{NCE}}\big(\theta;\,q,\,C_t(q),\,d^+\big) \;\ge\; \log(1 + e^z). \tag{3}$$

Step 2: Normalized softplus upper-bounds the indicator. $h(z) \coloneqq \log(1 + e^z)/\log 2$ is increasing, $h(0) = 1$, and $h(z) \ge 0$; thus

$$\mathbf{1}[z \ge 0] \;\le\; \frac{\log(1 + e^z)}{\log 2}. \tag{4}$$

Step 3: Combine Steps 1–2.

$$\begin{aligned}&\mathbf{1}\big[s_\theta(q,d^-) \ge s_\theta(q,d^+)\big]\\ &= \mathbf{1}[z \ge 0]\\ &\le \frac{\log(1 + e^z)}{\log 2}\\ &\le \frac{\mathcal{L}^{(1)}_{\mathrm{NCE}}(\theta; q, C_t(q), d^+)}{\log 2}.\end{aligned} \tag{5}$$

Averaging proves the claim.

A.2 Proof of Proposition 3.4

Let $Y = \mathbf{1}_{[g=+1]}$ be the clean anchor label on a flipped item ($\pi = \textsc{Yes}$). WF replaces Y by a constant weight $w \in [0,1]$. Consider

$$\begin{aligned}J(w) &= \mathbb{E}\big[(w - Y)^2 \mid \pi = \textsc{Yes}\big]\\ &= (w-1)^2 \;\Pr(Y{=}1 \mid \pi{=}\textsc{Yes}) + w^2 \;\Pr(Y{=}0 \mid \pi{=}\textsc{Yes}).\end{aligned} \tag{6}$$

Then $J'(w) = 2w - 2\,\Pr(Y{=}1 \mid \pi{=}\textsc{Yes})$, so the unique minimizer in $[0,1]$ is

$$w^\star = \Pr(Y{=}1 \mid \pi{=}\textsc{Yes}) = \Pr[g{=}+1 \mid \pi{=}\textsc{Yes}] = 1 - \sigma_t.$$

A.3 A bound on the per-anchor InfoNCE loss

Assume bounded logits $|s_\theta(q,\cdot)| \le B$ and $|C_t(q)| \le K$. For any anchor a (true or flipped),

$$\begin{aligned}\mathcal{L}^{(1)}(\theta; q, C_t(q), a) &= \log\!\Big(1 + \sum_{d \in C_t(q)} \exp\!\Big(\tfrac{s_\theta(q,d) - s_\theta(q,a)}{\tau}\Big)\Big)\\ &\le\; \log\!\big(1 + K\,e^{2B/\tau}\big) \;\eqqcolon\; \ell_{\max}(B, K, \tau).\end{aligned} \tag{7}$$

A.4 Proof of Proposition 3.3

Compare the *unweighted* clean objective (true positives only) to WF, which adds flipped anchors with weight w_{flip}. For a fixed $(q, C_t(q))$, write $F_t(q) = F_t^+(q) \,\dot\cup\, F_t^-(q)$ for truly positive vs. false flips. The per-list increment is

$$\Delta(q) = (w_{\mathrm{flip}} - 1) \sum_{a \in F_t^+(q)} \mathcal{L}^{(1)}(\theta; \cdot, a) \;+\; w_{\mathrm{flip}} \sum_{a \in F_t^-(q)} \mathcal{L}^{(1)}(\theta; \cdot, a). \tag{8}$$

Since $w_{\text{flip}} \le 1$, the first term is nonpositive; hence

$$\begin{aligned} \Delta(q)_+ \le\; & w_{\text{flip}} \sum_{a \in F_t^-(q)} \mathcal{L}^{(1)}(\theta; q, C_t(q), a) \\ \le\; & w_{\text{flip}}\, \ell_{\max}(B, K, \tau)\, |F_t^-(q)| \qquad \text{(using App. A.3).} \end{aligned} \tag{9}$$

Taking expectations and writing $m_t = \mathbb{E}_q[|F_t(q)|]$ and $\sigma_t = \Pr[g = -1 \mid a \in F_t(q)]$,

$$\zeta_t = \frac{1}{\log 2}\, \mathbb{E}[\Delta(q)_+] \;\le\; \underbrace{\frac{m_t}{\log 2}\, \ell_{\max}(B, K, \tau)}_{C_{\text{loss}}(B,K,\tau,m_t)}\; w_{\text{flip}}\, \sigma_t. \tag{10}$$

With $w_{\text{flip}}^{\star} = 1 - \sigma_t$, $\zeta_t \le C_{\text{loss}}\, \sigma_t(1 - \sigma_t) \le C_{\text{loss}}/4$.

A.5 Properties of $f(\rho)$ in Equation (5) to (6)

Setup and notation. Let $A \coloneqq 1 - \alpha > 0$ and $B \coloneqq \gamma \in [0, 1)$ with $A > B$ when $\alpha + \gamma < 1$. Fix a pool with *pre-judge* hidden-positive rate $\rho \in [0, 1]$.

Derivation of $f(\rho)$. Among the items the judge keeps as No,

$$\begin{aligned} f(\rho) &= \Pr(g = +1 \mid \pi = \text{No}) \\ &= \frac{\Pr(\pi = \text{No} \mid g = +1)\Pr(g = +1)}{\Pr(\pi = \text{No})} = \frac{B\rho}{A(1 - \rho) + B\rho}. \end{aligned} \tag{11}$$

Derivative at 0. f is smooth on $[0, 1)$ and

$$f'(\rho) = \frac{BA}{\big(A - (A - B)\rho\big)^2}; \qquad \text{hence} \quad f'(0) = \frac{B}{A} = \kappa.$$

Global quadratic upper bound. For all $\rho \in [0, 1]$,

$$f(\rho) \;\le\; \rho - \frac{A - B}{A}\, \rho(1 - \rho), \tag{12}$$

because

$$\rho - f(\rho) = \frac{A - B}{A - (A - B)\rho}\, \rho(1 - \rho) \;\ge\; \frac{A - B}{A}\, \rho(1 - \rho).$$

Local linear bound with explicit $\bar{\rho}(\epsilon)$. Using Taylor with remainder, $f(\rho) = \kappa\rho + \frac{\rho^2}{2} f''(\xi_\rho)$ where $f''(\rho) = \dfrac{2AB(A - B)}{\big(A - (A - B)\rho\big)^3}$. For $\rho \le A/(2(A - B))$, $f''(\rho) \le 16\, AB(A - B)/A^3$. Set

$$\bar{\rho}(\epsilon) \;\coloneqq\; \min\left\{ \frac{A}{2(A - B)},\; \frac{\epsilon A^3}{8AB(A - B)} \right\}.$$

Then $f(\rho) \le (\kappa + \epsilon)\rho$ for $\rho \in [0, \bar{\rho}(\epsilon)]$.

A.6 Proof of Lemma 3.6: Drifted recursion

Let (q, d) be drawn from the mixture $P_t^-(\cdot \mid q)$ after marginalizing q, and define the *global* hidden-positive rate $\rho_t \coloneqq \Pr[g = +1]$ under this mixture. Re-judge with the same (α, γ) to obtain U_t. By Bayes,

$$\begin{aligned} \mathbb{E}_{U_t} \mathbf{1}[g = +1] &= \Pr(g = +1 \mid \pi = \text{No}) \\ &= \frac{\gamma\, \rho_t}{(1 - \alpha)(1 - \rho_t) + \gamma\, \rho_t} = f(\rho_t). \end{aligned} \tag{13}$$

By the variational characterization of TV on $[0, 1]$,

$$\begin{aligned} \rho_{t+1} &= \mathbb{E}_{P_{t+1}^-} \mathbf{1}[g = +1] \\ &\le \mathbb{E}_{U_t} \mathbf{1}[g = +1] + \mathrm{TV}(P_{t+1}^-, U_t) = f(\rho_t) + \delta_t. \end{aligned} \tag{14}$$

Using the global quadratic bound on f yields

$$\rho_{t+1} \le \rho_t - c\, \rho_t(1 - \rho_t) + \delta_t \quad \text{with} \quad c = \tfrac{1 - \alpha - \gamma}{1 - \alpha}.$$

A.7 Proof of Lemma 3.7: Entry into the local region

Let $c = 1 - \kappa > 0$ and $g(\rho) = \rho - c\rho(1-\rho) + \bar{\delta}$. The fixed-point equation $g(\rho) = \rho$ is $c\rho(1-\rho) = \bar{\delta}$ with roots

$$\rho_{\pm} = \tfrac{1}{2}\Big(1 \pm \sqrt{1 - 4\bar{\delta}/c}\Big).$$

If $\rho_t \geq \frac{1}{2}$ then $\rho_{t+1} \leq \rho_t - \frac{c}{4} + \bar{\delta} < \rho_t$ since $\bar{\delta} < c/4$, so in finitely many steps $\rho_t \leq \frac{1}{2}$. If $\rho_t \in (\rho_-, \frac{1}{2}]$, then $c\rho_t(1-\rho_t) > \bar{\delta}$ and hence $\rho_{t+1} \leq g(\rho_t) < \rho_t$. Continuity implies finite-time entry into $[0, \rho_- + \eta] \subseteq [0, \bar{\rho}(\epsilon)]$ for some $\eta > 0$.

A.8 Proof of Theorem 3.8: Local geometric convergence

Inside the local region, $f(\rho) \leq r\rho$ so $\rho_{t+1} \leq r\rho_t + \delta_t$ for $t \geq T$. Unrolling,

$$\begin{aligned} \rho_t &\leq r^{t-T}\rho_T + \sum_{i=0}^{t-T-1} r^i\,\delta_{t-1-i} \\ &\leq r^{t-T}\rho_T + \frac{\bar{\delta}}{1-r}\,(1 - r^{t-T}). \end{aligned} \tag{15}$$

For pairwise risk,

$$\begin{aligned} X_t &= \Pr(g=+1)\,\mathbb{E}[I_t \mid g=+1] + \Pr(g=-1)\,\mathbb{E}[I_t \mid g=-1] + \zeta_t \\ &\leq \rho_t + (1-\rho_t)\,\eta + \zeta_t, \end{aligned} \tag{16}$$

using Lemma 3.2 and Assumption A2. Substitute Equation 23 to obtain Equation 8.

A.9 Proof of Proposition 3.9

From $X_t \leq \eta_t + (1-\eta_t)\rho_t + \zeta_t$ and $\eta_{t+1} \leq \eta_t$, $\zeta_{t+1} \leq \zeta_t$, we have

$$\begin{aligned} X_{t+1} - X_t &\leq (1-\eta_t)(\rho_{t+1} - \rho_t) + (\zeta_{t+1} - \zeta_t) \\ &\leq (1-\eta_t)\big(f(\rho_t) - \rho_t + \delta_t\big) + (\zeta_{t+1} - \zeta_t), \end{aligned} \tag{17}$$

using Lemma 3.6. Apply the quadratic bound on f for the sufficient condition.

A.10 Proof of Corollary 3.10

From $X_0 \leq \eta + (1-\eta)\rho_0 + \zeta_0$ and

$$X_\star = \eta + (1-\eta)\frac{\bar{\delta}}{1-r} + \bar{\zeta},$$

the condition

$$\rho_0 > \frac{\bar{\delta}}{1-r} + \frac{\bar{\zeta} - \zeta_0}{1-\eta}$$

implies $X_0 > X_\star$. Under forward correction $\zeta_0 = \bar{\zeta} = 0$.

A.11 Proof of Corollary 3.11

From (15), if $\delta_t \to 0$ then for any $\varepsilon > 0$ there exists M with $\sup_{j \geq M} \delta_j < \varepsilon$, so for all large t,

$$\sum_{i=0}^{t-T-1} r^i\,\delta_{t-1-i} \leq \varepsilon \sum_{i=0}^{\infty} r^i = \frac{\varepsilon}{1-r}.$$

Letting $t \to \infty$ gives $\limsup_t \rho_t \leq \varepsilon/(1-r)$; since ε is arbitrary, $\rho_t \to 0$. Then

$$X_t \leq \eta + (1-\eta)\rho_t + \zeta_t \to \eta; \qquad \text{if } \eta \to 0, \text{ then } X_t \to 0.$$

Sufficient conditions. By the drift decomposition in App. A.13, if the miner/judge operator is locally Lipschitz in θ with fixed abstention thresholds and θ_t stabilizes, then support overlap tends to one and in-support reweight drift vanishes, implying $\delta_t \to 0$. Under precision gating with $w_{\text{flip}} \leq 1 - \sigma_t$ and a judge whose $\alpha_t \downarrow 0$ at nontrivial prevalence $\pi_t^+ \in (0,1)$, Eq.(2) gives $\sigma_t \to 0$ and hence $\zeta_t \leq C_{\text{loss}}\sigma_t(1-\sigma_t) \to 0$ by Prop.3.3 and Prop. 3.4.

A.12 Forward correction is unbiased

Proposition A.1 (Forward correction is unbiased). *Under class-conditional noise with confusion matrix T, the forward-corrected listwise loss ℓ_{fc} satisfies $\mathbb{E}_{\pi|g}[\ell_{\text{fc}}] = \ell_{\text{clean}}$ for each $(q, C_t(q), d^+)$; hence $\mathbb{E}[\ell_{\text{fc}}] = \mathbb{E}[\ell_{\text{clean}}]$.*

Sketch. Let ℓ_{clean} be the clean per-list loss and ℓ_{fc} the forward-corrected one. With class-conditional noise $\Pr(\pi = k \mid g = j) = T_{jk}$, the forward correction replaces the observed one-hot over π by T^{-1} times the observed label vector. Linearity gives $\mathbb{E}_{\pi|g}[\ell_{\text{fc}}] = \ell_{\text{clean}}$, hence $\mathbb{E}[\ell_{\text{fc}}] = \mathbb{E}[\ell_{\text{clean}}]$.

A.13 Drift bound via pool churn

Definition A.2 (Total variation (TV) drift budget). Let $U_t(\cdot \mid q)$ be the distribution obtained by re-judging draws from $P_t^-(\cdot \mid q)$ with the same (α, γ). Define

$$\delta_t \;\coloneqq\; \mathbb{E}_q\big[\mathrm{TV}\big(P_{t+1}^-(\cdot \mid q),\, U_t(\cdot \mid q)\big)\big]. \tag{18}$$

Two precise bounds. We provide (i) a general bound that separates support churn and in-intersection reweight drift, and (ii) a cardinality corollary under a uniform-within-support mining model.

General bound (support churn + reweight drift). For each q, let $M_t(\cdot \mid q)$ be the mining distribution on $\mathcal{D}$ and $\mathcal{J}$ the judge-and-gating operator mapping μ to the post-judge "No" distribution

$$\mathcal{J}(\mu)(A) \;\coloneqq\; \frac{\displaystyle\int_A \mathbf{1}\{\pi(q,d) = \text{No}\}\, d\mu(d)}{\displaystyle\int_{\mathcal{D}} \mathbf{1}\{\pi(q,d) = \text{No}\}\, d\mu(d)}. \tag{19}$$

Then $P_t^-(\cdot \mid q) = \mathcal{J}(M_t(\cdot \mid q))$ and $U_t(\cdot \mid q) = \mathcal{J}(P_t^-(\cdot \mid q))$. By triangle inequality,

$$\begin{aligned}\mathrm{TV}\big(P_{t+1}^-(\cdot \mid q), U_t(\cdot \mid q)\big) \;&\le\; \mathrm{TV}\big(\mathcal{J}(M_{t+1}),\, \mathcal{J}(M_t)\big)\\ &\quad+\; \underbrace{\mathrm{TV}\big(\mathcal{J}(M_t),\, \mathcal{J}(P_t^-)\big)}_{=0},\end{aligned} \tag{20}$$

since $\mathcal{J}$ is idempotent on its image. Decompose M_t and M_{t+1} by the intersection $I_t(q) = \operatorname{supp} M_t \cap \operatorname{supp} M_{t+1}$ and its complement to obtain

$$\mathrm{TV}\big(\mathcal{J}(M_{t+1}), \mathcal{J}(M_t)\big) \;\le\; |\lambda_{t+1} - \lambda_t| \;+\; \lambda_\star\, \mathrm{TV}\big(\mathcal{J}(\widetilde{M}_{t+1}^I), \mathcal{J}(\widetilde{M}_t^I)\big), \tag{21}$$

where $\lambda_t = M_t(I_t(q))$, $\lambda_{t+1} = M_{t+1}(I_t(q))$, $\lambda_\star = \max\{\lambda_t, \lambda_{t+1}\}$. As $\mathcal{J}$ is 1-Lipschitz in TV when restricted to a fixed support (policy fixed on $I_t(q)$), let

$$\omega_t(q) \coloneqq \mathrm{TV}\big(\widetilde{M}_{t+1}^I(\cdot \mid q),\, \widetilde{M}_t^I(\cdot \mid q)\big), \qquad \chi_t^{(\text{mass})} \coloneqq \mathbb{E}_q\big[|\lambda_{t+1}(q) - \lambda_t(q)|\big],$$

we get

$$\delta_t \;\le\; \chi_t^{(\text{mass})} \;+\; \Omega_t, \qquad \Omega_t \coloneqq \mathbb{E}_q[\omega_t(q)].$$

Cardinality corollary (uniform within support). If the miner samples *uniformly* on finite supports $S_t(q)$ and $S_{t+1}(q)$ (e.g., top-K lists), and the judge policy is fixed on the intersection, then

$$\chi_t^{(\text{mass})} \;\le\; \mathbb{E}_q\Big[1 - \frac{|S_{t+1}(q) \cap S_t(q)|}{|S_t(q)|}\Big] \;\eqqcolon\; \chi_t, \qquad \Omega_t = 0,$$

hence $\delta_t \le \chi_t$. Approximate ANN/top-K effects contribute additively by a small ζ_t^{ANN}, giving $\delta_t \le \chi_t + \zeta_t^{\text{ANN}}$.

A.14 Derivations for Equation (2) and (4)

By Bayes,

$$\begin{aligned}\sigma_t &= \Pr[g = -1 \mid \pi = \textsc{Yes}]\\ &= \frac{\Pr(\pi = \textsc{Yes} \mid g = -1)\Pr(g = -1)}{\Pr(\pi = \textsc{Yes})}\\ &= \frac{\alpha(1 - \pi_t^+)}{(1 - \gamma)\pi_t^+ + \alpha(1 - \pi_t^+)},\end{aligned} \tag{22}$$

giving Eq.(2) . To enforce $\sigma_t \le \sigma^\star \in (0, 1)$, solve for α:

$$\alpha(1 - \pi_t^+) \le \sigma^\star\big((1 - \gamma)\pi_t^+ + \alpha(1 - \pi_t^+)\big) \quad\Longleftrightarrow\quad \alpha \le \frac{\sigma^\star(1 - \gamma)\pi_t^+}{(1 - \pi_t^+)(1 - \sigma^\star)}, \tag{23}$$

which is Eq.(4).

Lemma A.3 (Minimax-safe flip weighting). *Let $\sigma_t = \Pr[g = -1 \mid \pi = \textsc{Yes}]$ and suppose the per-anchor InfoNCE loss is bounded by $\ell_{\max}$ (App. A.3). Then for*

$$G_+(w) \coloneqq \big(\mathbb{E}[\widetilde{\mathcal{L}}_{\text{WF}}^{(1)}] - \mathbb{E}[\mathcal{L}_{\text{clean}}^{(1)}]\big)_+, \quad G_-(w) \coloneqq \big(\mathbb{E}[\mathcal{L}_{\text{clean}}^{(1)}] - \mathbb{E}[\widetilde{\mathcal{L}}_{\text{WF}}^{(1)}]\big)_+, \tag{24}$$

we have $G_+(w) \le \ell_{\max} m_t w \sigma_t$ and $G_-(w) \le \ell_{\max} m_t (1 - w)(1 - \sigma_t)$; the minimax $w^\star = \arg\min_w \max\{G_+(w), G_-(w)\}$ equals $1 - \sigma_t$. Moreover, for any $w \le 1 - \sigma_t$,

$$\max\{G_+(w), G_-(w)\} \;\le\; \ell_{\max}\, m_t\, (1 - w)\, (1 - \sigma_t) \;\le\; \ell_{\max}\, m_t\, \sigma_t(1 - \sigma_t),$$

with equality at $w = 1 - \sigma_t$.

B Proofs for Section 4

B.1 Proof of Proposition 4.1 (bias under fixed adversary)

Let $F(\theta, \tau) = \nabla_\theta L_{\text{distill}}(\theta; \phi^\star) + \tau \nabla_\theta L_{\text{adv}}(\theta; \phi^\star)$ with $\tau = \lambda_{\text{adv}}/\lambda_{\text{distill}}$. Since $L_{\text{distill}}(\cdot; \phi^\star)$ is C^2 and locally strongly convex at $\theta^\star$, we have $F(\theta^\star, 0) = 0$ and $\partial_\theta F(\theta^\star, 0) = H_{\text{distill}} \succ 0$. By the Implicit Function Theorem there exists a C^1 curve $\theta(\tau)$ with $F(\theta(\tau), \tau) = 0$, $\theta(0) = \theta^\star$, and $\theta'(0) = -H_{\text{distill}}^{-1} g_{\text{adv}}$ where $g_{\text{adv}} = \nabla_\theta L_{\text{adv}}(\theta^\star; \phi^\star)$. Hence $\hat{\theta} = \theta(\tau) = \theta^\star - \tau H_{\text{distill}}^{-1} g_{\text{adv}} + O(\tau^2)$.

B.2 Proof of Proposition 4.2 (variance inflation)

Let unbiased mini-batch gradients be $\widehat{g}_{\text{distill}}$ and $\widehat{g}_{\text{adv}}$ with covariances Σ_{distill} and Σ_{adv}. For $\widehat{g} = \lambda_{\text{distill}} \widehat{g}_{\text{distill}} + \lambda_{\text{adv}} \widehat{g}_{\text{adv}}$,

$$\operatorname{Var}[\widehat{g}] = \lambda_{\text{distill}}^2 \Sigma_{\text{distill}} + \lambda_{\text{adv}}^2 \Sigma_{\text{adv}} + \lambda_{\text{distill}} \lambda_{\text{adv}} (\Sigma_\times + \Sigma_\times^\top).$$

Near $\theta^\star$, continuity implies $\operatorname{tr} \operatorname{Var}[\widehat{g}] > \lambda_{\text{distill}}^2 \operatorname{tr} \Sigma_{\text{distill}}$ for any fixed $\lambda_{\text{adv}} > 0$ as long as Σ_{adv} is nonzero, proving strict inflation.

B.3 Proof of Theorem 4.3 (convergence with ALD)

Consider SGD

$$\theta_{t+1} = \theta_t - \eta_t \Big(\nabla L_{\text{distill}}(\theta_t; \phi_t) + \lambda_{\text{adv}}(t) \nabla L_{\text{adv}}(\theta_t; \phi_t) + \xi_t \Big),$$

with $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$, $\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] \le C$, $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, and bounded iterates. Add and subtract $\nabla L_{\text{distill}}(\theta_t; \phi^\star)$:

$$\begin{aligned} \theta_{t+1} = \theta_t - \eta_t \Big(& \nabla L_{\text{distill}}(\theta_t; \phi^\star) \\ & + \underbrace{\Delta_t^{(\phi)}}_{\nabla L_{\text{distill}}(\theta_t; \phi_t) - \nabla L_{\text{distill}}(\theta_t; \phi^\star)} \\ & + \underbrace{\lambda_{\text{adv}}(t) \nabla L_{\text{adv}}(\theta_t; \phi_t)}_{\Delta_t^{(\text{adv})}} \\ & + \xi_t \Big). \end{aligned}$$

If $\phi_t \to \phi^\star$ and gradients are locally Lipschitz, then $\|\Delta_t^{(\phi)}\| \to 0$; ALD imposes $\lambda_{\text{adv}}(t) \to 0$ and $\sum_t \eta_t \lambda_{\text{adv}}(t) < \infty$, so $\sum_t \eta_t \|\Delta_t^{(\text{adv})}\| < \infty$ (bounded gradients locally). Thus the recursion is a Robbins–Monro scheme for the limiting ODE $\dot{\theta} = -\nabla L_{\text{distill}}(\theta; \phi^\star)$ with a summable perturbation and square-summable noise. Standard stochastic approximation results imply almost-sure convergence to the stationary set of $L_{\text{distill}}(\cdot; \phi^\star)$; the bias therefore vanishes and the adversarial variance contribution decays as $O(\lambda_{\text{adv}}(t)^2)$.