Appendix

A Proof of Proposition 3.1

Proposition 3.1. *Assume g is β-smooth, twice differentiable, and locally α-strongly convex in w around $\{w_{T-K-1}, \dots, w_T\}$. Let $\Xi_{t+1}(w_t, \lambda) = w_t - \gamma \nabla_w g(w_t, \lambda)$. For $\gamma \leq \frac{1}{\beta}$, it holds*

$$\|h_{T-K} - \mathrm{d}_\lambda f\| \leq 2^{T-K+1}(1-\gamma\alpha)^K \|\nabla_{\hat{w}^*} f\| M_B \tag{9}$$

where $M_B = \max_{t\in\{0,\dots,T-K\}} \|B_t\|$. In particular, if g is globally α-strongly convex, then

$$\|h_{T-K} - \mathrm{d}_\lambda f\| \leq \tfrac{(1-\gamma\alpha)^K}{\gamma\alpha} \|\nabla_{\hat{w}^*} f\| M_B. \tag{10}$$

Proof. Let $\mathrm{d}_\lambda f - h_{T-K} = e_K$. By definition of h_{T-K},

$$e_K = \left(\sum_{t=0}^{T-K} B_t A_{t+1} \cdots A_{T-K} \right) A_{T-K+1} \cdots A_T \nabla_{\hat{w}^*} f$$

Therefore, when g is locally α-strongly convex with respect to w in the neighborhood of $\{w_{T-K-1}, \dots, w_T\}$,

$$\begin{aligned} \|e_K\| &\leq \| \sum_{t=0}^{T-K} B_t A_{t+1} \cdots A_{T-K} \| \| A_{T-K+1} \cdots A_T \nabla_{\hat{w}^*} f \| \\ &\leq (1-\gamma\alpha)^K \|\nabla_{\hat{w}^*} f\| \| \sum_{t=0}^{T-K} B_t A_{t+1} \cdots A_{T-K} \| \end{aligned}$$

Suppose g is β-smooth but nonconvex. In the worst case, if the smallest eigenvalue of $\nabla_{w,w} g(w_{t-1}, \lambda)$ is $-\beta$, then $\|A_t\| = 1 + \gamma\beta \leq 2$ for $t = 0, \dots, T-K$. This gives the bound in (9). However, if g is globally strongly convex, then

$$\|e_K\| \leq \|\nabla_{\hat{w}^*} f\| (1-\gamma\alpha)^K \max_{t\in\{0,\dots,T-K\}} \|B_t\| \sum_{t=0}^{T-K} (1-\gamma\alpha)^t$$

The bound (10) uses the fact that $\sum_{t=0}^{T-K} (1-\gamma\alpha)^t \leq \sum_{t=0}^{\infty} (1-\gamma\alpha)^t = \frac{1}{\gamma\alpha}$ ■

B Proof of Lemma 3.2

Lemma 3.2. *Let g be globally strongly convex and $\nabla_\lambda f = 0$. Assume g is second-order continuously differentiable and B_t has full column rank for all t. Let $\Xi_{t+1}(w_t, \lambda) = w_t - \gamma \nabla_w g(w_t, \lambda)$. For all $K \geq 1$, with T large enough and γ small enough, there exists $c > 0$, s.t. $h_{T-K}^\top \mathrm{d}_\lambda f \geq c \|\nabla_{\hat{w}^*} f\|^2$. This implies h_{T-K} is a sufficient descent direction, i.e. $h_{T-K}^\top \mathrm{d}_\lambda f \geq \Omega(\|\mathrm{d}_\lambda f\|^2)$.*

Proof. To illustrate the idea, here we prove the case where $K = 1$. For $K > 1$, similar steps can be applied. To prove the statement, we first expand the inner product by definition

$$h_{T-1}^\top \mathrm{d}_\lambda f = \|h_{T-1}\|^2 + (B_T \nabla_{\hat{w}^*} f)^\top \left(\sum_{t=0}^{T-1} B_t A_{t+1} \cdots A_{T-1} \right) A_T \nabla_{\hat{w}^*} f$$

where we recall $h_{T-1} = B_T \nabla_{\hat{w}^*} f$ as $\nabla_\lambda f = 0$ by assumption.

Next we show a technical lemma, which provides a critical tool to bound the second term above; its proof is given in the next section.

Lemma B.1. *Let g be α-strongly convex and β-smooth. Assume B_t and A_t are Lipschitz continuous in w, and assume B_T has full column rank. For $\gamma \leq \frac{1}{\beta}$,*

$$\begin{aligned}&\left(B_T \nabla_{\hat{w}^*} f\right)^\top B_t A_{t+1} \cdots A_T \nabla_{\hat{w}^*} f \\ &\geq (1-\gamma\alpha)^{T-t} \|B_T \nabla_{\hat{w}^*} f\|^2 - \|\nabla_{\hat{w}^*} f\|^2 O\left(\frac{e^{-\alpha\gamma(T-1)}}{1-e^{-\alpha\gamma}} + (\gamma(\beta-\alpha))^{T-t}\right)\end{aligned}$$

By Lemma B.1, we can then write

$$h_{T-1}^\top \mathrm{d}_\lambda f \geq \|B_T \nabla_{\hat{w}^*} f\|^2 \left(1 + \sum_{t=0}^{T-1} (1-\gamma\alpha)^{T-t}\right) - \|\nabla_{\hat{w}^*} f\|^2 O\left(\sum_{t=0}^{T-1} \frac{e^{-\alpha\gamma(T-1)}}{1-e^{-\alpha\gamma}} + (\gamma(\beta-\alpha))^{T-t}\right)$$

Because

$$\sum_{t=0}^{T-1} (\gamma(\beta-\alpha))^{T-t} = \sum_{k=1}^{T} (\gamma(\beta-\alpha))^k \leq \frac{\gamma(\beta-\alpha)}{1-\gamma(\beta-\alpha)} \qquad (\because \gamma \leq \beta)$$

and $B_T^\top B_T$ is non-singular by assumption,

$$\begin{aligned}h_{T-1}^\top \mathrm{d}_\lambda f &\geq \|\nabla_{\hat{w}^*} f\|^2 \Omega(1) - \|\nabla_{\hat{w}^*} f\|^2 O\left(\frac{T e^{-\alpha\gamma(T-1)}}{1-e^{-\alpha\gamma}} + \frac{\gamma(\beta-\alpha)}{1-\gamma(\beta-\alpha)}\right) \\ &\geq C \|\nabla_{\hat{w}^*} f\|^2\end{aligned}$$

for some $c > 0$, when T is large enough and γ is small enough. The implication holds because $\|\mathrm{d}_\lambda f\| \leq O(\|\nabla_{\hat{w}^*} f\|)$. ■

B.1 Proof of Lemma B.1

Proof. Let C_A and C_B be the Lipschitz constant of A_t and B_t. First, we see that the inner product can be lower bounded by the following terms

$$\left(B_T \nabla_{\hat{w}^*} f\right)^\top B_t A_{t+1} \cdots A_T \nabla_{\hat{w}^*} f \geq (1-\gamma\alpha)^{T-t} \|B_T \nabla_{\hat{w}^*} f\|^2 - \Delta_1 - \Delta_2 - \Delta_3$$

where

$$\Delta_1 = C_B \|B_T \nabla_{\hat{w}^*} f\| \|\nabla_{\hat{w}^*} f\| \|w_{T-1} - w_{t-1}\| \|A_{t+1} \cdots A_T\|$$

$$\Delta_2 = C_A \|B_T^\top B_T \nabla_{\hat{w}^*} f\| \|\nabla_{\hat{w}^*} f\| \sum_{k=t+1}^{T-1} \|w_{T-1} - w_{k-1}\| \|A_{t+1} \cdots A_{k-1}\| \|A_T\|^{T-k}$$

$$\Delta_3 = \|\nabla_{\hat{w}^*} f\| \|B_T^\top B_T \nabla_{\hat{w}^*} f\| \|A_k - (1-\gamma\alpha) I\|^{T-k}$$

The above lower bounds can be shown by the following inequalities:

$$\begin{aligned}&\left(B_T \nabla_{\hat{w}^*} f\right)^\top B_t A_{t+1} \cdots A_T \nabla_{\hat{w}^*} f \\ &\geq \nabla_{\hat{w}^*} f^\top \left(B_T^\top B_T\right) A_{t+1} \cdots A_T \nabla_{\hat{w}^*} f - C_B \|B_T \nabla_{\hat{w}^*} f\| \|w_{T-1} - w_{t-1}\| \|A_{t+1} \cdots A_T \nabla_{\hat{w}^*} f\|\end{aligned}$$

$$\begin{aligned}&\nabla_{\hat{w}^*} f^\top \left(B_T^\top B_T\right) A_{t+1} \cdots A_T \nabla_{\hat{w}^*} f \\ &\geq \nabla_{\hat{w}^*} f^\top \left(B_T^\top B_T\right) A_{t+1} \cdots A_{T-2} A_T^2 \nabla_{\hat{w}^*} f - C_A \|w_{T-1} - w_{T-2}\| \|A_{t+1} \cdots A_{T-2}\| \|A_T\| \|B_T^\top B_T \nabla_{\hat{w}^*} f\| \|\nabla_{\hat{w}^*} f\| \\ &\geq \nabla_{\hat{w}^*} f^\top B_T^\top B_t A_T^{T-t} \nabla_{\hat{w}^*} f - C_A \|B_T^\top B_T \nabla_{\hat{w}^*} f\| \|\nabla_{\hat{w}^*} f\| \sum_{k=t+1}^{T-1} \|w_{T-1} - w_{k-1}\| \|A_{t+1} \cdots A_{k-1}\| \|A_T\|^{T-k}\end{aligned}$$

$$\nabla_{\hat{w}^*} f^\top B_T^\top B_T A_T^{T-t} \nabla_{\hat{w}^*} f \geq (1-\gamma\alpha)^{T-t} \nabla_{\hat{w}^*} f^\top B_T^\top B_T \nabla_{\hat{w}^*} f - \|\nabla_{\hat{w}^*} f\| \|B_T^\top B_T \nabla_{\hat{w}^*} f\| \|A_T - (1-\gamma\alpha) I\|^{T-t}$$

Next we upper bound the error terms: Δ_1, Δ_2, and Δ_3. We will use the fact that gradient descent converges linearly when optimizing a strongly convex and smooth function [25].

Lemma B.2. *Let w_0 be the initial condition. Running gradient descent to optimize an α-strongly convex and β-smooth function g, with step size $0 < \gamma \leq \frac{1}{\beta}$, generates a sequence $\{w_t\}$ satisfying*

$$\|w_t - w^*\| \leq De^{-\alpha\gamma t} \tag{13}$$

where $D = \|w_0 - w^*\|$ and $w^* = \arg\min g(w)$.

Lemma B.2 implies for $T \geq t$, $\|w_T - w_t\| \leq 2De^{-\alpha\gamma t}$.

Now we proceed to bound the errors Δ_1, Δ_2, and Δ_3.

Bound on Δ_1 Because

$$\begin{aligned}\|w_{T-1} - w_{t-1}\|\|A_{t+1}\cdots A_T\| &\leq 2De^{-\alpha\gamma(t-1)}(1-\gamma\alpha)^{T-t}\\ &\leq 2De^{-\alpha\gamma(t-1)}e^{-\gamma\alpha(T-t)}\\ &= 2De^{-\alpha\gamma(T-1)}\end{aligned}$$

we can upper bound Δ_1 by

$$\begin{aligned}\Delta_1 &= C_B\|B_T\nabla_{\hat{w}^*}f\|\|\nabla_{\hat{w}^*}f\|\|w_{T-1} - w_{t-1}\|\|A_{t+1}\cdots A_T\|\\ &\leq \|B_T\nabla_{\hat{w}^*}f\|\|\nabla_{\hat{w}^*}f\| \times 2C_BDe^{-\alpha\gamma(T-1)}\end{aligned}$$

Bound on Δ_2 Because

$$\begin{aligned}\sum_{k=t+1}^{T-1}\|w_{T-1} - w_{k-1}\|\|A_{t+1}\cdots A_{k-1}\|\|A_T\|^{T-k} &\leq \sum_{k=t+1}^{T-1} 2De^{-\alpha\gamma(k-1)}(1-\alpha\gamma)^{k-1-t+T-k}\\ &\leq 2D(1-\alpha\gamma)^{T-t-1}\sum_{k=t+1}^{T-1}e^{-\alpha\gamma(k-1)}\\ &\leq 2D(1-\alpha\gamma)^{T-t-1}e^{-\alpha\gamma t}\sum_{k=t+1}^{T-1}e^{-\alpha\gamma(k-t-1)}\\ &\leq 2De^{-\alpha\gamma(T-1)}\sum_{m=0}^{T-t}e^{-\alpha\gamma m}\\ &\leq \frac{2D}{1-e^{-\alpha\gamma}}e^{-\alpha\gamma(T-1)}\end{aligned}$$

we can upper bound Δ_2 by

$$\begin{aligned}\Delta_2 &= C_A\|B_T^\top B_T\nabla_{\hat{w}^*}f\|\|\nabla_{\hat{w}^*}f\|\sum_{k=t+1}^{T-1}\|w_{T-1} - w_{k-1}\|\|A_{t+1}\cdots A_{k-1}\|\|A_T\|^{T-k}\\ &= \|B_T^\top B_T\nabla_{\hat{w}^*}f\|\|\nabla_{\hat{w}^*}f\| \times \frac{2C_AD}{1-e^{-\alpha\gamma}}e^{-\alpha\gamma(T-1)}\end{aligned}$$

Bound on Δ_3 Because

$$\|A_k - (1-\gamma\alpha)I\| = \|\gamma\left(\alpha I - \nabla_w^2 f(w_{k-1})\right)\| \leq \gamma(\beta - \alpha)$$

we can upper bound Δ_3 by

$$\Delta_3 = \|\nabla_{\hat{w}^*}f\|\|B_T^\top B_T\nabla_{\hat{w}^*}f\|\|A_t - (1-\gamma\alpha)I\|^{T-t} \leq \|\nabla_{\hat{w}^*}f\|\|B_T^\top B_T\nabla_{\hat{w}^*}f\|\left(\gamma(\beta-\alpha)\right)^{T-t}$$

Final Result Using the bounds on Δ_1, Δ_2, and Δ_3, we prove the final result.

$$\begin{aligned}
&\left(B_T \nabla_{\hat{w}^*} f\right)^\top B_t A_{t+1} \cdots A_T \nabla_{\hat{w}^*} f \\
&\geq (1-\gamma\alpha)^{T-t} \|B_T \nabla_{\hat{w}^*} f\|^2 - \Delta_1 - \Delta_2 - \Delta_3 \\
&\geq (1-\gamma\alpha)^{T-t} \|B_T \nabla_{\hat{w}^*} f\|^2 - \|\nabla_{\hat{w}^*} f\|^2 O\left(\frac{e^{-\alpha\gamma(T-1)}}{1-e^{-\alpha\gamma}} + (\gamma(\beta-\alpha))^{T-t}\right)
\end{aligned}$$

because B_T has full column rank and

$$\begin{aligned}
\Delta_1 + \Delta_2 + \Delta_3 &\leq \|\nabla_{\hat{w}^*} f\|^2 \left(\frac{2C_A D}{1-e^{-\alpha\gamma}} e^{-\alpha\gamma(T-1)} + 2C_B D e^{-\alpha\gamma(T-1)} + (\gamma(\beta-\alpha))^{T-k}\right) \\
&= \|\nabla_{\hat{w}^*} f\|^2 \times O\left(\frac{e^{-\alpha\gamma(T-1)}}{1-e^{-\alpha\gamma}} + (\gamma(\beta-\alpha))^{T-t}\right)
\end{aligned}$$

■

C Proof of Theorem 3.3

Theorem 3.3. *Suppose F is smooth and bounded below, and suppose there is $\epsilon < \infty$ such that $\|h_{T-K} - \mathrm{d}_\lambda f\| \leq \epsilon$. Using h_{T-K} as a stochastic first-order oracle with a decaying step size $\eta_\tau = O(1/\sqrt{\tau})$ to update λ with gradient descent, it follows after R iterations,*

$$\mathbb{E}\left[\sum_{\tau=1}^{R} \frac{\eta_\tau \|\nabla F(\lambda_\tau)\|^2}{\sum_{\tau=1}^{R} \eta_\tau}\right] \leq \widetilde{O}\left(\epsilon + \frac{\epsilon^2+1}{\sqrt{R}}\right).$$

That is, under the assumptions in Proposition 3.1, learning with h_{T-K} converges to an ϵ-approximate stationary point, where $\epsilon = O((1-\gamma\alpha)^{-K})$.

Proof. The proof of this theorem is a standard proof of non-convex optimization with biased gradient estimates. Here we include it for completeness, as part of it will be used later in the proof of Theorem 3.4.

Let λ_τ be the τth iterate. For short hand, we write $\mathrm{d}_\lambda f_{(\tau)} = \mathrm{d}_\lambda f(\lambda_\tau)$, and $h_{T-K,(\tau)} = h_{T-K}(\lambda_\tau)$. Assume F is L-smooth and $\|\mathrm{d}_\lambda f_{(\tau)}\| \leq G$ and $\|h_{T-K,(\tau)}\| \leq G$ almost surely for all τ. Then by L-smoothness, it satisfies

$$F(\lambda_{\tau+1}) \leq F(\lambda_\tau) + \langle \nabla F(\lambda_\tau), \lambda_{\tau+1} - \lambda_\tau \rangle + \frac{L}{2}\|\lambda_{\tau+1} - \lambda_\tau\|^2.$$

Let $e_\tau = \mathrm{d}_\lambda f_{(\tau)} - h_{T-K,(\tau)}$ be the error in the gradient estimate. Substitute the recursive update $\lambda_{\tau+1} = \lambda_\tau - \eta_t h_{T-K,(\tau)}$ to the above inequality. Conditioned on λ_τ, it satisfies

$$\mathbb{E}_{|\lambda_\tau}[F(\lambda_{\tau+1})] \leq F(\lambda_\tau) + \mathbb{E}_{|\lambda_\tau}\left[-\eta_t \langle \nabla F(\lambda_\tau), h_{T-K,(\tau)} \rangle + \frac{L\eta_t^2}{2}\|h_{T-K,(\tau)}\|^2\right].$$

Because

$$\begin{aligned}
-\mathbb{E}_{|\lambda_\tau}[\langle \nabla F(\lambda_\tau), h_{T-K,(\tau)} \rangle] &= \mathbb{E}_{|\lambda_\tau}\left[-\langle \nabla F(\lambda_\tau), \mathrm{d}_\lambda f_{(\tau)} \rangle + \langle \nabla F(\lambda_\tau), e_\tau \rangle\right] \\
&\leq -\|\nabla F(\lambda_\tau)\|^2 + G\|e_\tau\| \qquad (14)
\end{aligned}$$

and

$$\frac{1}{2}\|h_{T-K,(\tau)}\|^2 = \frac{1}{2}\|\mathrm{d}_\lambda f_{(\tau)}\|^2 + \frac{1}{2}\|e_\tau\|^2 - \langle \mathrm{d}_\lambda f_{(\tau)}, h_{T-K,(\tau)} \rangle \leq \frac{3G^2}{2} + \frac{1}{2}\|e_\tau\|^2$$

we can upper bound $\mathbb{E}_{|\lambda_\tau}[F(\lambda_{\tau+1})]$ as

$$\mathbb{E}_{|\lambda_\tau}[F(\lambda_{\tau+1})] \leq F(\lambda_\tau) + \mathbb{E}_{|\lambda_\tau}\left[-\eta_\tau \|\nabla F(\lambda_\tau)\|^2 + \eta_\tau G\|e_\tau\| + L\eta_\tau^2\left(\frac{3G^2}{2} + \frac{1}{2}\|e_\tau\|^2\right)\right]$$

Performing telescoping sum with the above inequality, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{\tau=1}^{R}\eta_\tau\|\nabla F(\lambda_\tau)\|^2\right] &\le F(\lambda_1)+\mathbb{E}\left[\sum_{\tau=1}^{R}G\eta_\tau\|e_\tau\|+L\eta_\tau^2\left(\frac{3G^2}{2}+\frac{1}{2}\|e_\tau\|^2\right)\right]\\ &\le F(\lambda_1)+\sum_{\tau=1}^{R}\left(G\epsilon\eta_\tau+\frac{L(3G^2+\epsilon^2)}{2}\eta_\tau^2\right)\end{aligned}$$

Dividing both sides by $\sum_{\tau=1}^{R}\eta_\tau$ and using the facts that $\eta_\tau = O(\frac{1}{\sqrt{\tau}})$ and that

$$\frac{\sum_{\tau=1}^{R}\frac{1}{\tau}}{\sum_{\tau=1}^{R}\frac{1}{\sqrt{\tau}}} = O\left(\frac{\log R}{\sqrt{R}}\right)$$

proves the theorem. ■

D Proof of Theorem 3.4

Theorem 3.4. *Under the assumptions in Proposition 3.1 and Theorem 3.3, if in addition*

1. *g is second-order continuously differentiable*
2. *B_t has full column rank around w_T*
3. *$\nabla_\lambda f^\top(\mathrm{d}_\lambda f + h_{T-K} - \nabla_\lambda f) \ge \Omega(\|\nabla_\lambda f\|^2)$*
4. *the problem is deterministic (i.e. $F = f$)*

then for all $K \ge 1$, with T large enough and γ small enough, the limit point is an exact stationary point, i.e. $\lim_{\tau\to\infty}\|\nabla F(\lambda_\tau)\| = 0$.

Proof. First we consider the special case when S is deterministic. Let $H \ge K$. We decompose the full gradients into four parts

$$\nabla F = \mathrm{d}_\lambda f = \nabla_\lambda f + q + r + e$$

where

$$\begin{aligned}q &= \sum_{t=T-K+1}^{T} B_t A_{t+1}\cdots A_T \nabla_{\hat{w}^*} f\\ r &= \sum_{t=T-H+1}^{T-K} B_t A_{t+1}\cdots A_T \nabla_{\hat{w}^*} f\\ e &= \sum_{t=0}^{T-H} B_t A_{t+1}\cdots A_T \nabla_{\hat{w}^*} f\end{aligned}$$

We assume that w_t enters a locally strongly convex region for $t \ge H$. This implies, by Proposition 3.1, that $\|e\| \le O(e^{-\alpha\gamma H}\|\nabla_{\hat{w}^*} f\|)$.

To prove the theorem, we first verify two conditions:

1. By Lemma 3.2, the assumption $\nabla_\lambda f^\top(\mathrm{d}_\lambda f + h_{T-K} - \nabla_\lambda f) \ge \Omega(\|\nabla_\lambda f\|^2)$, and $\|e\| \le O(e^{-\alpha\gamma H}\|\nabla_{\hat{w}^*} f\|)$:

$$\begin{aligned}\mathrm{d}_\lambda f^\top h_{T-K} &= (\nabla_\lambda f + q + r + e)^\top(\nabla_\lambda f + q)\\ &= \|\nabla_\lambda f\|^2 + \nabla_\lambda f^\top(q+e+r) + q^\top\nabla_\lambda f + q^\top(q+r) + q^\top e\\ &\ge \Omega(\|\nabla_\lambda f\|^2) + q^\top(q+r) + q^\top e && \text{(Assumption)}\\ &\ge \Omega(\|\nabla_\lambda f\|^2) + \Omega(\|\nabla_{\hat{w}^*} f\|^2) + q^\top e && \text{(Lemma 3.2)}\\ &\ge \Omega(\|\nabla_\lambda f\|^2) + \Omega(\|\nabla_{\hat{w}^*} f\|^2) - O\left(e^{-\alpha\gamma H}\|\nabla_{\hat{w}^*} f\|^2\right) && (\|e\| \le O(e^{-\alpha\gamma H}\|\nabla_{\hat{w}^*} f\|))\end{aligned}$$

where we note

$$\begin{aligned} \mathrm{d}_\lambda f + h_{T-K} - \nabla_\lambda f &= \nabla_\lambda f + q + r + e + \nabla_\lambda f + q - \nabla_\lambda f \\ &= \nabla_\lambda f + q + r + e + q \end{aligned}$$

Therefore, for H large enough, it holds that

$$\mathrm{d}_\lambda f^\top h_{T-K} \geq \Omega(\|\nabla_\lambda f\|^2 + \|\nabla_{\hat{w}^*} f\|^2) \tag{15}$$

2. By definition of $h_{T-K} = \nabla_\lambda f + q$, it holds that

$$\|h_{T-K}\|^2 \leq 2\|\nabla_\lambda f\|^2 + 2\|q\|^2 \leq O(\|\nabla_\lambda f\|^2 + \|\nabla_{\hat{w}^*} f\|^2) \tag{16}$$

Next, we prove a lemma

Lemma D.1. *Let f be a lower-bound and L-smooth function. Consider the iterative update rule*

$$x_{t+1} = x_t - \eta g_t$$

where g_t satisfies $g_t^\top \nabla f(x_t) \geq c_1 h_t^2$ and $\|g_t\|^2 \leq c_2 h_t^2$, for some constant $c_1, c_2 > 0$ and scalar h_t. Suppose f is lower-bounded and η is chosen such that $\left(-c_1\eta + \frac{Lc_2\eta^2}{2}\right) \leq 0$. Then $\lim_{t\to\infty} h_t = 0$.

Proof. By L-smoothness,

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= -\eta \nabla f(x_t)^\top g_t + \frac{L\eta^2}{2}\|g_t\|^2 \\ &\leq \left(-c_1\eta + \frac{Lc_2\eta^2}{2}\right) h_t^2 \end{aligned}$$

By telescoping sum, we can show $\sum_{t=0}^{\infty} \left(c\eta - \frac{L\eta^2}{2}\right) h_t^2 < \infty$, which implies $\lim_{t\to\infty} h_t = 0$. ■

Finally, we prove the main theorem by applying Lemma D.1. Consider a deterministic problem. Take $h_t^2 = \|\nabla_\lambda f(\lambda_t)\|^2 + \|\nabla_{\hat{w}^*} f(\lambda_t)\|^2$. Because of (15) and (16), by Lemma D.1, it satisfies that

$$\lim_{t\to\infty} h_t = \lim_{t\to\infty} \|\nabla_\lambda f(\lambda_t)\|^2 + \|\nabla_{\hat{w}^*} f(\lambda_t)\|^2 = 0$$

As $\|\mathrm{d}_\lambda f\| \leq O(\|\nabla_\lambda f\| + \|\nabla_{\hat{w}^*} f\|)$, it shows $\|\mathrm{d}_\lambda f\|$ converges to zero in the limit.

■

E Proof of Theorem 3.5

Theorem 3.5. *There is a problem, satisfying all but assumption 3 in Theorem 3.4, such that optimizing λ with h_{T-K} does not converge to a stationary point.*

Proof. We prove the non-convergence using the following strategy. First we show that, when assumption 3 in Theorem 3.4, i.e.

$$\nabla_\lambda f^\top (\mathrm{d}_\lambda f + h_{T-K} - \nabla_\lambda f) \geq \Omega(\|\nabla_\lambda f\|^2) \tag{17}$$

does not hold, there is some problem such that $h_{T-k} \neq 0$ for all stationary points (i.e. λ such that $\mathrm{d}_\lambda f = 0$). Then we show that, for such a problem, optimizing λ with h_{T-k} cannot converge to any of the stationary points.

Counter example To construct the counterexample, we consider a scalar deterministic bilevel optimization problem of the form

$$\begin{aligned} &\min_{\lambda} \frac{1}{2}(\hat{w}^*)^2 + \phi(\lambda) \\ &\text{s.t.} \quad \hat{w}^* \approx w^* \in \arg\min_{w} \frac{1}{2}(w-\lambda)^2 \end{aligned} \tag{18}$$

in which ϕ is some perturbation function that we will later define, and $\hat{w}^*$ is computed by performing $T > 1$ steps of gradient descent in the lower-level optimization problem with some constant initial condition w_0 and constant step size $0 < \gamma < 1$, i.e.

$$\hat{w}^* = w_T, \qquad w_{t+1} = w_t - \gamma(w_t - \lambda)$$

We can observe this problem satisfies *almost* all the assumptions in Theorem 3.4:

1. The lower-level objective g is smooth and strongly convex. (Proposition 3.1)
2. The upper-level objective F is smooth. (Theorem 3.3)
3. The lower-level objective g is second-order continuously differentiable (assumption 1 in Theorem 3.4)
4. The Jacobian if full rank, i.e. $B_t = \gamma > 0$ (assumption 2 in Theorem 3.4)
5. The upper-level objective function is deterministic, i.e. $F = f$ (assumption 4 in Theorem 3.4)

But we will show that properly setting ϕ can break the non-interfering assumption in (17) (i.e. assumption 3 in Theorem 3.4) and then creates a problem such that optimizing λ with K-RMD does not converge to an exact stationary point.

We follow the two-step strategy mentioned above.

Step 1: Non-vanishing approximate gradient Without loss of generality, let us consider optimizing λ with 1-RMD. In this case we can write the approximate and the exact gradients in closed form as

$$h_{T-1} = \nabla\phi + w^*\gamma, \qquad \mathrm{d}_\lambda f = \nabla\phi + w^*\gamma \sum_{t=0}^{T} (1-\gamma)^{T-t} \tag{19}$$

which are given by (5) and (8). We will show that by properly choosing ϕ, we can define $f(\lambda) = \frac{1}{2}(\hat{w}^*)^2 + \phi(\lambda)$ such that, at any of the stationary points of f, the approximate gradient of 1-RMD does not vanish. That is, we show when $\mathrm{d}_\lambda f = 0$, $h_{T-1} \neq 0$.

Before proceeding, let us define $u = w^*\gamma$ and $v = w^*\gamma \sum_{t=0}^{T}(1-\gamma)^{T-t}$ for convenience. To show how to construct ϕ, let us consider the stationary points in the case[9] when $\phi = 0$. Let P_0 denote the set of these stationary points, i.e. $P_0 = \{\lambda : v = 0\}$. Since f is smooth and lower-bounded, we know that P_0 is non-empty, and from the construction of our counterexample we know that P_0 contains exactly the λs such that $w^* = 0$.

This implies that for $\lambda \in \mathbb{R} \backslash P_0$, it satisfies $w^* \neq 0$ and therefore

$$uv = (w^*\gamma)^2 \sum_{t=0}^{T} (1-\gamma)^{T-t} > 0 \tag{20}$$

We use this fact to pick an adversarial ϕ. Consider any smooth, lower-bounded ϕ whose stationary points are not in P_0, e.g. $\phi(\lambda) = \frac{1}{2}(\lambda - \lambda_0)^2$ and $\lambda_0 \notin P_0$. Then $f(\lambda) = \frac{1}{2}(\hat{w}^*)^2 + \phi(\lambda)$ has a non-empty set of stationary points

[9] Note in this special case, assumption 3 in Theorem 3.4 holds trivially when $\phi(\lambda) = 0$ (i.e. $\nabla_\lambda f = 0$) and optimizing λ with K-RMD converges to an exact stationary point.

P_ϕ such that $P_\phi \cap P_0 = \emptyset$. We see that, for such ϕ, the non-interfering assumption (assumption 3 in Theorem 3.4) is violated in P_ϕ:

$$\begin{aligned}
\nabla_\lambda f^\top (\mathrm{d}_\lambda f + h_{T-1} - \nabla_\lambda f) &= \nabla_\lambda f^\top (\nabla_\lambda f + u - \nabla_\lambda f) && \because \mathrm{d}_\lambda f = 0 \text{ and } h_{T-1} = \nabla_\lambda f + u \\
&= \nabla_\lambda \phi^\top u \\
&= -vu && \because 0 = \mathrm{d}_\lambda f = \nabla_\lambda \phi + v \\
&< 0 && \because (20) \text{ and } P_\phi \cap P_0 = \emptyset \\
&< (\nabla_\lambda \phi)^2 && \because v > 0 \text{ for } \lambda \in P_\phi
\end{aligned}$$

And we show for any $\lambda \in P_\phi$ it holds that $h_{T-1} \neq 0$. This can be seen from the definition

$$h_{T-1} = \nabla \phi + u = \mathrm{d}_\lambda f + u - v = u - v \neq 0$$

where the last inequality is because $w^* \neq 0$ for $\lambda \in P_\phi$.

Step 2: Non-convergence to any stationary point We have shown that there is a problem which satisfies all the assumptions but assumption 3 of Theorem 3.4, and at any of its stationary points (i.e. when $\mathrm{d}_\lambda f = 0$) we have $h_{T-K} \neq 0$. Now we show this property implies failure to converge to the stationary points for the general problems considered in Theorem 3.5 (i.e. we do not rely on the form made in Step 1 anymore).

We prove this by contradiction. Let λ^* be one of the stationary points. We choose $\delta_0 > 0$ such that, for some $\epsilon > 0$, $\|h_{T-K}\| > \epsilon/\gamma$ for all λ inside the neighborhood $\{\lambda : \|\lambda - \lambda^*\| < \frac{\delta_0}{2}\}$, where we recall γ is the step size of the lower-level optimization problem. A non-zero δ_0 exists because h_{T-1} is continuous by our assumption and $h_{T-K} \neq 0$ at λ^*.

We are ready to show the contradiction. Let $\delta = \min\{\delta_0, \epsilon\}$. Suppose there is a sequence $\{\lambda_\tau\}$ that converges to the stationary point λ^*. This means that there is $0 < M < \infty$ such that, $\forall \tau \geq M$, $\|\lambda_\tau - \lambda^*\| < \frac{\delta}{2}$, which implies that $\forall \tau \geq M$, $\|\lambda_{\tau+1} - \lambda_\tau\| < \delta$. However, by our choice of δ_0, $\|\lambda_{\tau+1} - \lambda_\tau\| = \gamma \|h_{T-K}\| > \epsilon \geq \delta$, leading to a contradiction.

Thus, no sequence $\{\lambda_\tau\}$ converges to any of the stationary points. This concludes our proof. ■

F Proof of Proposition 3.6

Proposition 3.6. *Under the assumptions in Proposition 3.1, suppose w_t converges to a stationary point w^*. Let $A_\infty = \lim_{t\to\infty} A_t$ and $B_\infty = \lim_{t\to\infty} B_t$. For $\gamma < \frac{1}{\beta}$, it satisfies that*

$$-\nabla_{\lambda,w} g \nabla_{w,w}^{-1} g = B_\infty \textstyle\sum_{k=0}^{\infty} A_\infty^k \tag{12}$$

Proof. Recall our shorthand that $\nabla_{\lambda,w} g$ and $\nabla_{w,w} g$ are evaluated at (w^*, λ). In the limit, it holds that

$$\begin{aligned}
\lim_t A_t &= \lim_t \nabla_w \Xi_t(w_{t-1}, \lambda) = \nabla_w (w^* - \gamma \nabla_w g(w^*, \lambda)) = I - \gamma \nabla_{w,w} g =: A_\infty \\
\lim_t B_t &= \lim_t \nabla_\lambda \Xi_t(w_{t-1}, \lambda) = \nabla_\lambda (w^* - \gamma \nabla_w g(w^*, \lambda)) = -\gamma \nabla_{\lambda,w} g =: B_\infty
\end{aligned}$$

To prove the equality (12), we use Lemma (F.1).

Lemma F.1. *[32] For a matrix A with $\|A\| < 1$, it satisfies that*

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$$

Since $\gamma \leq \frac{1}{\beta}$, we have $\gamma\alpha I \preceq \gamma \nabla_{w,w} g \preceq I$, so $\|I - \gamma \nabla_{w,w} g\| < 1$. By Lemma F.1,

$$\nabla_{w,w}^{-1} g = \gamma (I - I + \gamma \nabla_{w,w} g)^{-1} = \gamma \sum_{k=0}^{\infty} (I - \gamma \nabla_{w,w} g)^k = \gamma \sum_{k=0}^{\infty} A_\infty^k$$

Therefore,

$$-\nabla_{\lambda,w} g \nabla_{w,w}^{-1} g = (-\gamma \nabla_{\lambda,w} g) \left(\frac{1}{\gamma} \nabla_{w,w}^{-1} g \right) = B_\infty \sum_{k=0}^{\infty} A_\infty^k$$

■

G Detailed experimental setup

In this appendix, we provide more details about the settings we used in each experiment. We use Adam [33] to optimize the upper-level objective and vanilla gradient descent for the lower objective. We denote by $\hat{w}^*$ the results of running T steps of gradient descent with step size γ.

G.1 Data hypercleaning

In this appendix, we provide more details about the data hypercleaning experiment on MNIST from Section 4.2.1.

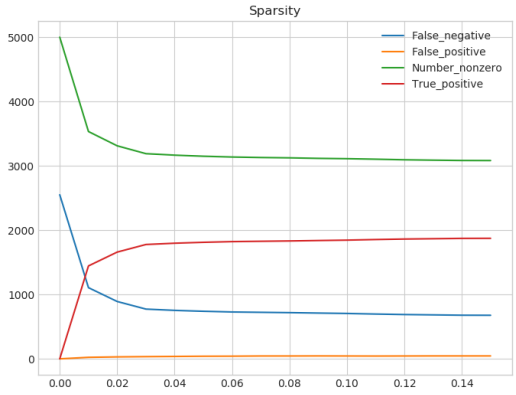
Both the training and the validation sets consist of 5000 class-balanced examples from the MNIST dataset. The test set consists of the remaining examples. For each training example, with probability $\frac{1}{2}$, we replaced the label with a uniformly random one.

For various K, we performed K-RMD for 1000 hyperiterations. Like in the toy experiment (Section 4.1) we adjusted the initial meta-learning rate η_0 for each K so that the norm of the initial update was roughly the same for each K.

We asserted earlier that the reported F1 scores are not sensitive to our choice of threshold $\lambda_i < -3$. To validate this assertion, we repeated the experiment for various thresholds. F1 scores are reported in the table below.

K	$\lambda_i < -4$	$\lambda_i < -3$	$\lambda_i < -1$
1	0.84	0.84	0.84
5	0.89	0.89	0.90
25	0.89	0.89	0.89
50	0.89	0.89	0.89
100	0.89	0.89	0.89

We only ran these experiments for 150 hyperiterations, because the F1 score has essentially converged by that point. Indeed, the plot below shows identification of corrupted labels for $K = 1$, with cutoff $\lambda_i < -4$. The X axis is in units of 1000 hyperiterations. We see that 1-RMD rapidly identifies most of the mislabeled examples, with a few false positives.



G.2 Task interaction

We use $T = 100$ iterations of gradient descent with learning rate 0.1 in the lower objective which yields $\hat{w}_S^*$. To ensure that C is symmetric, and that C_{ij} and ρ are nonnegative, we re-parametrize them as $\rho = \text{softplus}(\nu)$ and

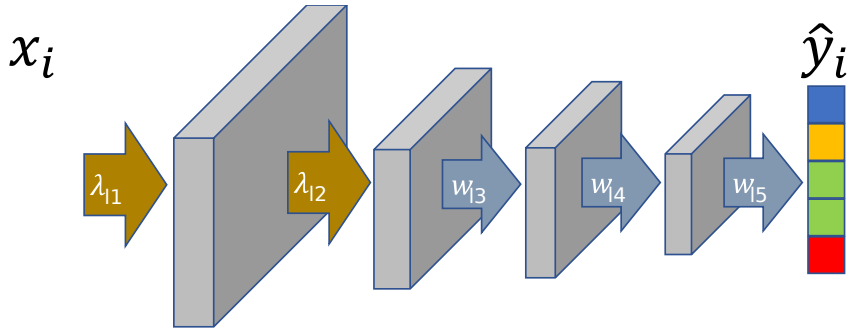


Figure 8: One-shot learning network architecture. The first two convolutional layers map the input image into a "hyper-representation" space which is frozen while optimizing the lower-level objective. The last three layers are tuned for each task and regularized to avoid overfitting. All the convolutional layers have 64 3×3 kernels. There is a max-pooling layer followed by a batch-normalization and a ReLU layer after each convolution.

$C = A + A^\top$, where $A_{ij} = \text{softplus}(B_{ij})$ and B is a hyperparameter matrix. Thus, the hyperparameters to be optimized are $\lambda = \{B, \nu\}$.

Rather than using raw pixels, we extract image features from the output of the average pooling layer in Resnet-18 [34] which is trained on ImageNet [35]. We use the same data pre-processing that is used for training Resnet architecture.

When reporting test accuracy, we run 10 independent trials. In each trial, we sample the training and validation datasets with a balanced set of m examples each ($m = 50$ for CIFAR-10 and $m = 300$ for CIFAR-100) and use the rest of the dataset for testing. To avoid over-fitting, we use early stopping when the testing error does not improve for 500 hyper-iterations.

Although we are using a similar setting as Franceschi et al. [9], our results on full back-propagation are quite different from theirs. We believe it is because we are using a different network architecture and pre-processing method for feature extraction.

G.3 One-shot classification

Dataset The Omniglot dataset [31], a popular benchmark for few-shot learning, is used in this experiment. We consider 5-way classification with 1 training and 15 validation examples for each of the five classes. To evaluate the generalization performance, we restrict the meta-training dataset to a random subset of 1200 of the 1623 Omniglot characters. The meta-validation dataset consists of 100 other characters, and meta-testing dataset has the remaining 323 characters. We use the meta-validation dataset for tuning the upper-level optimization parameters and report the performance of the algorithm on the meta-testing dataset. Note that no data augmentation method is used in the training.

Neural Network and Optimization The overall neural network architecture is shown in Figure 8. Our architecture inherits the hyper-representation model of Franceschi et al. [2] with some modifications. The first two convolutional layers, parametrized by hyperparameter $\lambda = \{\lambda_{l_1}, \lambda_{l_2}\}$, transform the input image into a "hyper-representation" space. The last three layers, parametrized by $w = \{w_{l_3}, w_{l_4}, w_{l_5}\}$ are fine-tuned in the lower-level optimization. Additionally, we have regularization hyperparameters $\lambda_r = \{\rho_i\}_{i=1}^3 \cup \{c_j\}_{j=1}^3$. The overall setup corresponds essentially to meta-learning the two bottom layers of a CNN; for each task, the weights in the first two layers are frozen, and the k-way classifier of the last three layers is fine tuned. Overall, the model has $\approx$ 110k hyperparameters and $\approx$ 75k parameters.

We use a meta-batch-size of 4 in each hyper-iteration. To limit the training time, we stop all the algorithms after 5000 hyper-iterations. Needless to say, these results could be further improved by using data augmentation, higher meta-batch size, and running more hyper-iterations. However, our current setup is selected so that all the experiments can be run in a reasonable amount of time, while sharing a similar setting used in practical one-shot learning.