A. Appendix

A.1. Symbols and Notations of C-Flat Turbo

- model parameter: $\boldsymbol{\theta}$;
- SAM perturbed model parameter: $\boldsymbol{\theta}+\boldsymbol{\epsilon}_0^* = \boldsymbol{\theta}+\rho\cdot\frac{\nabla\mathcal{L}(\boldsymbol{\theta})}{\|\nabla\mathcal{L}(\boldsymbol{\theta})\|}$;
- proxy model parameter: $\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^* = \boldsymbol{\theta}+\rho\cdot(\boldsymbol{g}_s-\boldsymbol{g})/\|\boldsymbol{g}_s-\boldsymbol{g}\|$;
- proxy perturbed model parameter: $\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*+\rho\cdot\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*)/\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*)\|$;
- the empirical loss: $\mathcal{L}(\boldsymbol{\theta})$, with its gradient $\boldsymbol{g}$;
- the SAM loss: $\mathcal{L}_{SAM}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta})+\mathcal{R}_\rho^0(\boldsymbol{\theta}) = \max_{\|\boldsymbol{\epsilon}_0\|\le\rho}\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_0)$ with its gradient $\boldsymbol{g}_s$;
- the C-Flat loss: $\mathcal{L}_{CFlat}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta})+\mathcal{R}_\rho^0(\boldsymbol{\theta})+\lambda\cdot\mathcal{R}_\rho^1(\boldsymbol{\theta}) = \max_{\|\boldsymbol{\epsilon}_0\|\le\rho}\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_0)+\lambda\cdot\rho\max_{\|\boldsymbol{\epsilon}_1\|\le\rho}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1)\right\|$, with its gradient $\boldsymbol{g}_s+\lambda\boldsymbol{g}_f$;
- the gradient of proxy model: $\boldsymbol{g}_0 = \nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*)$;
- the gradient of proxy perturbed model: $\boldsymbol{g}_1 = \nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*+\rho\cdot\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*)/\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*)\|)$;
- the empirical loss term: $\boldsymbol{g} = \nabla\mathcal{L}(\boldsymbol{\theta})$;
- the zeroth-order sharpness term: $\boldsymbol{g}_s-\boldsymbol{g} = \nabla\mathcal{R}_\rho^0(\boldsymbol{\theta})$;
- the first-order flatness term: $\boldsymbol{g}_f = \nabla\mathcal{R}_\rho^1(\boldsymbol{\theta})$;
- the direction-invariant sharpness component: $\boldsymbol{g}_{vs} = \boldsymbol{g}_s-\frac{\langle\boldsymbol{g}_s,\boldsymbol{g}\rangle}{\|\boldsymbol{g}\|^2}\boldsymbol{g}$;
- the direction-invariant flatness component: $\boldsymbol{g}_{vf} = \boldsymbol{g}_f-\frac{\langle\boldsymbol{g}_f,\boldsymbol{g}_0\rangle}{\|\boldsymbol{g}_0\|^2}\boldsymbol{g}_0$.

A.2. Derivation of Equation 5

Following [6, 61], the gradient of the first-order flatness loss $\mathcal{R}_\rho^1$ is:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{R}_\rho^1(\boldsymbol{\theta}) &= \rho\cdot\nabla_{\boldsymbol{\theta}}\max_{\boldsymbol{\epsilon}\in B(0,\rho)}\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon})\| \\ &= \rho\cdot\nabla_{\boldsymbol{\theta}}\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*)\| \\ &= \rho\cdot\left(\frac{\partial}{\partial\boldsymbol{\theta}}\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*)\|+\frac{\partial\boldsymbol{\epsilon}_1^*}{\partial\boldsymbol{\theta}}\cdot\nabla_{\boldsymbol{\epsilon}}\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon})\|\Big|_{\boldsymbol{\epsilon}=\boldsymbol{\epsilon}_1^*}\right) \\ &\approx \rho\cdot\nabla_{\boldsymbol{\theta}}\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*)\|\end{aligned} \tag{10}$$

Here, ϵ_1^* denotes the optimal perturbation that maximizes the gradient norm within the ℓ_2-ball $B(0,\rho)$. To make the computation tractable, we approximate it using first-order Taylor expansion and finite differences:

$$\begin{aligned}\boldsymbol{\epsilon}_1^* &= \arg\max_{\boldsymbol{\epsilon}\in B(0,\rho)}\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon})\| \\ &\approx \arg\max_{\boldsymbol{\epsilon}\in B(0,\rho)}\left(\left(\nabla_{\boldsymbol{\theta}}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|\right)^T\boldsymbol{\epsilon}\right) \\ &= \rho\cdot\frac{\nabla_{\boldsymbol{\theta}}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|}{\|\nabla_{\boldsymbol{\theta}}\|\nabla\mathcal{L}(\boldsymbol{\theta})\|\|} \\ &\approx \rho\cdot\frac{\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\delta})-\nabla\mathcal{L}(\boldsymbol{\theta})}{\|\nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\delta})-\nabla\mathcal{L}(\boldsymbol{\theta})\|} \\ &= \rho\cdot\frac{\boldsymbol{g}_s-\boldsymbol{g}}{\|\boldsymbol{g}_s-\boldsymbol{g}\|},\end{aligned} \tag{11}$$

where $\boldsymbol{g} = \nabla\mathcal{L}(\boldsymbol{\theta})$, $\boldsymbol{g}_s = \nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\delta})$, and $\boldsymbol{\delta} = \rho'\cdot\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|}$ is a small perturbation in the direction of the gradient, with $\rho'\ll\rho$.

Let $\boldsymbol{\theta}_p = \boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*$ be the perturbed model after maximizing the gradient norm. Then Eq. 10 continues as:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{R}_\rho^1(\boldsymbol{\theta}) &\approx \rho\cdot\nabla_{\boldsymbol{\theta}}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_p)\right\| \\ &\approx \frac{\rho}{\rho'}\cdot\left[\nabla\mathcal{L}\left(\boldsymbol{\theta}_p+\rho'\cdot\frac{\nabla\mathcal{L}(\boldsymbol{\theta}_p)}{\|\nabla\mathcal{L}(\boldsymbol{\theta}_p)\|}\right)-\nabla\mathcal{L}(\boldsymbol{\theta}_p)\right],\end{aligned} \tag{12}$$

where $\rho'\ll\rho$ is a small step size for finite-difference approximation.

In our symbol table, we define

$$\boldsymbol{g}_0 := \nabla\mathcal{L}(\boldsymbol{\theta}_p) = \nabla\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon}_1^*), \tag{13}$$

$$\boldsymbol{g}_1 := \nabla\mathcal{L}\left(\boldsymbol{\theta}_p+\rho'\cdot\frac{\nabla\mathcal{L}(\boldsymbol{\theta}_p)}{\|\nabla\mathcal{L}(\boldsymbol{\theta}_p)\|}\right), \tag{14}$$

Then Eq. 12 can be written as the finite-difference form

$$\boldsymbol{g}_f \approx \frac{\rho}{\rho'}(\boldsymbol{g}_1-\boldsymbol{g}_0). \tag{15}$$

For simplicity of notation, we absorb the constant factor ρ/ρ' into λ, so that at the level of directions we can view

$$\boldsymbol{g}_f \propto \boldsymbol{g}_1-\boldsymbol{g}_0.$$

Finally, we define the direction-invariant component of $\boldsymbol{g}_f$ with respect to $\boldsymbol{g}_0$:

$$\boldsymbol{g}_{vf} = \boldsymbol{g}_f-\frac{\langle\boldsymbol{g}_f,\boldsymbol{g}_0\rangle}{\|\boldsymbol{g}_0\|^2}\cdot\boldsymbol{g}_0, \tag{16}$$

which is orthogonal to $\boldsymbol{g}_0$ and is the quantity tracked by the EMA variable in the implementation.

A.3. Hyperparameter Settings

We report the main hyperparameter settings for methods trained on CIFAR100 in Table 4. For the other datasets, we follow the original settings from the open-source repository and keep $k = 5$ and $\beta = 0.8$ fixed to balance efficiency and performance.

We also study the sensitivity of the scheduler and trigger threshold in C-Flat Turbo on CIFAR100. For the scheduler $k_t = k_0+c\cdot\frac{n}{N}$, larger k_0 generally improves throughput with negligible performance changes, while $c = 10$ provides a good trade-off between speed and accuracy, as shown in Table 5. This suggests that moderately increasing the turbo interval over the task sequence is sufficient to obtain most of the efficiency gains. For the trigger threshold $\|\boldsymbol{g}_{0j}\|^2 > \mu_{f,j}+m\cdot\sigma_{f,j}$, Table 6 shows that m between 0.2 and 1 performs best overall. Larger values improve efficiency but tend to degrade performance, whereas smaller values may lead to less stable trigger behavior.

Table 4. Hyperparameter settings for CIFAR100.

Methods	Epochs	LR	BS	Tasks	Exemplar	ρ	λ	k	β
iCaRL	20	1e-3	32	10	20	0.1	0.2	5	0.8
MEMO	20	1e-3	32	10	20	0.1	0.2	5	0.8
L2P	5	2e-3	16	10	-	0.02	0.2	5	0.8
Ranpac	5	1e-2	16	10	-	0.05	0.2	5	0.8
EASE	5	2.5e-3	16	10	-	0.05	0.2	5	0.8

Table 5. Sensitivity analysis of the scheduler hyperparameters in C-Flat Turbo on CIFAR100.

Optimizer	C-Flat Turbo								
k_0	2	2	2	5	5	5	10	10	10
c	5	10	20	5	10	20	5	10	20
Avg	92.19	92.18	92.05	92.08	92.08	92.01	92.07	92.04	92.04
Last	87.67	87.59	87.45	87.62	87.54	87.42	87.51	87.59	87.46
Img/s	99.17	99.91	63.19	104.86	102.74	99.05	104.64	106.89	107.39

Table 6. Sensitivity analysis of the trigger threshold hyperparameter in C-Flat Turbo on CIFAR100.

m	0.1	0.2	0.5	1	2	5
Avg	91.98	92.07	92.05	92.06	91.86	91.94
Last	87.43	87.48	87.39	87.50	87.07	87.16
Img/s	75.56	105.43	97.69	102.74	154.52	135.18

A.4. Memory Usage

The cached gradients are used to substitute partial sharpness-aware gradient computations, so their memory usage heavily depends on the number of trainable parameters in the model. As shown in Table 7, although larger models require more cached gradients, the overall memory overhead remains almost negligible relative to the expansion typically caused by large architectures.

Table 7. Memory usage of different architectures. EASE uses a frozen backbone, while iCaRL updates the full backbone.

Method	Optimizer	Backbone	Trainable / total params	Memory
EASE	C-Flat	ViT-Base-16	1.19M / 86.99M	2.14GB
	C-Flat Turbo	ViT-Base-16	1.19M / 86.99M	2.15GB
	C-Flat	ViT-Large-16	3.17M / 306.47M	5.34GB
	C-Flat Turbo	ViT-Large-16	3.17M / 306.47M	5.37GB
iCaRL	C-Flat	ResNet-18	11.17M / 11.17M	1.55GB
	C-Flat Turbo	ResNet-18	11.17M / 11.17M	1.66GB
	C-Flat	ResNet-34	21.28M / 21.28M	2.32GB
	C-Flat Turbo	ResNet-34	21.28M / 21.28M	2.51GB

A.5. Comparison to Similar Works

We further compare C-Flat Turbo with SS-SAM and AE-SAM on L2P and EASE, as shown in Table 8. Overall, C-Flat Turbo consistently achieves the best performance on both benchmarks, while maintaining competitive efficiency.

Table 8. Comparison with SS-SAM and AE-SAM on L2P (left) and EASE (right).

	L2P			EASE		
Method	Avg	Last	Img/s	Avg	Last	Img/s
+SS-SAM	89.50	84.45	77.42 (70.20%)	90.90	87.87	91.65 (54.99%)
+AE-SAM	89.70	84.48	80.86 (73.32%)	92.03	87.37	112.58 (67.54%)
+C-Flat Turbo	**89.78**	**84.69**	65.50 (59.4%)	**92.36**	**87.96**	102.74 (61.6%)

A.6. Other CL Settings

We further evaluate C-Flat Turbo with DUCT on DIL and with DualPrompt on TIL and CIL. As shown in Tables 9 and 10, Turbo consistently improves performance over the corresponding baselines.

Table 9. Results on DomainNet under the DIL setting.

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg
DUCT	74.21	60.52	67.73	63.50	69.51	68.63	67.35
+C-Flat	**74.93**	60.88	**68.10**	63.30	70.22	69.41	67.81
+C-Flat Turbo	74.89	**60.91**	68.01	**63.75**	**70.82**	**69.50**	**67.98**

Table 10. Results on CIFAR100 under the CIL and TIL settings.

Method	CIL		TIL	
	Avg	Last	Avg	Last
DualPrompt	85.77	**89.10**	97.79	**99.40**
+C-Flat	85.85	87.80	97.67	99.30
+C-Flat Turbo	**86.01**	87.00	**97.94**	**99.40**

A.7. Per-task Accuracy and Ablation Studies

Per-task accuracy provides a more detailed view of the continual learning process. As shown in Table 11, the reuse mechanism significantly reduces training time with minimal performance loss, while the linear step-size scheduler further improves speed, particularly for longer task sequences. The adaptive trigger additionally accelerates training, as it allows basic single propagation gradient descent in certain stages.

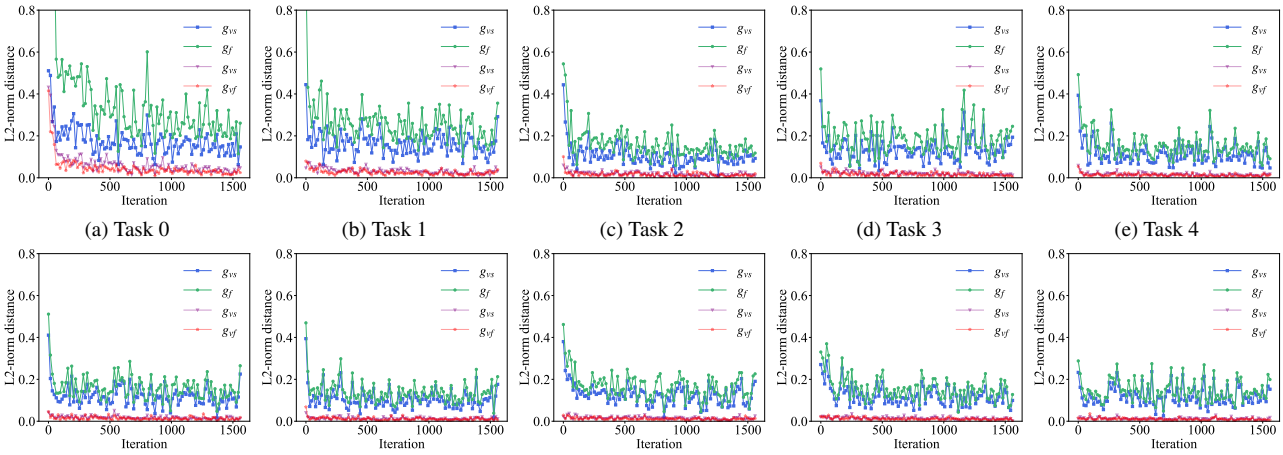
Regarding performance gains, prior works have shown that selectively applying SAM updates can outperform applying SAM throughout training. For instance, SS-SAM [63] explicitly demonstrates that with appropriate scheduling, models can achieve comparable or even superior performance at substantially lower computational cost compared to training exclusively with SAM. Similar observations also have been reported for AE-SAM [24] and SAM-In-Late-Phase [72].

A.8. Detailed Evolution of Gradient Distances

Figure 7 shows the L2-norm distances between sharpness and flatness gradients and their reference gradients across

Table 11. Per-task accuracy and ablation study results for EASE trained on the 10-split CIFAR100 dataset.

Method	reuse	scheduler	trigger	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Avg	Img/s
EASE	×	×	×	98.40	96.25	94.63	93.88	91.80	90.92	90.47	88.09	87.58	87.17	91.92	166.67
+C-Flat	×	×	×	98.50	96.45	94.87	94.08	91.94	91.05	90.64	88.44	87.93	87.58	92.15	44.25
	✓	×	×	98.50	96.37	94.77	93.94	91.92	91.05	90.67	88.28	87.81	87.45	92.08	67.20
+C-Flat Turbo	✓	✓	×	98.40	96.31	94.74	93.89	91.90	91.00	90.70	88.25	87.75	87.40	92.03	74.63
	✓	✓	✓	98.50	**96.60**	**95.07**	**94.15**	**92.08**	**91.27**	**90.73**	**88.52**	**88.00**	**87.57**	**92.25**	**102.74**



(a) Task 0 (b) Task 1 (c) Task 2 (d) Task 3 (e) Task 4

(f) Task 5 (g) Task 6 (h) Task 7 (i) Task 8 (j) Task 9

Figure 7. Visualization of L2-norm distances of the gradients every 5 steps across 10 tasks.

tasks. While g and g_0 exhibit significant fluctuations during training, the gradients g_{vs} and g_{vf}, core to zeroth-order sharpness and first-order flatness regularization, change much more slowly. This stability suggests their potential as shortcut directions for flat region exploration, bypassing the need for model ascent and backpropagation.

A.9. Convergence of Turbo in the Surrogate Steps

With the definitions in Appendix A.1, we denote C-Flat gradient at iteration t as

$$\boldsymbol{G}_t := \nabla \mathcal{L}_{\text{CFlat}}(\boldsymbol{\theta}_t) = \boldsymbol{g}_s + \lambda \boldsymbol{g}_f, \tag{17}$$

then we can rewrite

$$\boldsymbol{G}_t = \underbrace{\boldsymbol{g}}_{\text{empirical gradient}} + \underbrace{(\boldsymbol{g}_s - \boldsymbol{g})}_{\text{sharpness increment}} + \lambda \underbrace{\boldsymbol{g}_f}_{\text{flatness term}}, \tag{18}$$

where $\boldsymbol{g} = \nabla \mathcal{L}(\boldsymbol{\theta}_t)$ is the empirical gradient.

In the exact C-Flat updates, the sharpness increment $(\boldsymbol{g}_s - \boldsymbol{g})$ is obtained by subtracting the empirical gradient from the SAM gradient, and the flatness term $\boldsymbol{g}_f$ is obtained via the finite-difference approximation $\boldsymbol{g}_f \approx \frac{\rho}{\rho'}(\boldsymbol{g}_1 - \boldsymbol{g}_0)$ as in Eq. 15.

In Turbo, we do not recompute $\boldsymbol{g}_s$ and $\boldsymbol{g}_f$ at every surrogate step. Instead, we maintain EMA states $\boldsymbol{g}_{vs,t}$ and $\boldsymbol{g}_{vf,t}$ to track

$$\boldsymbol{u}_t^{(s)} := \boldsymbol{g}_s - \boldsymbol{g}, \qquad \boldsymbol{u}_t^{(f)} := \boldsymbol{g}_f,$$

and we construct deterministic surrogates

$$A_t(\boldsymbol{g}_{vs,t-1}) \approx \boldsymbol{u}_t^{(s)}, \tag{19}$$

$$B_t(\boldsymbol{g}_{vf,t-1}) \approx \boldsymbol{u}_t^{(f)}. \tag{20}$$

Therefore, at a Turbo surrogate step, the update direction is modeled as

$$\tilde{\boldsymbol{g}}_t = \underbrace{\boldsymbol{g}}_{\nabla \mathcal{L}(\boldsymbol{\theta}_t)} + A_t(\boldsymbol{g}_{vs,t-1}) + \lambda B_t(\boldsymbol{g}_{vf,t-1}) + \boldsymbol{\xi}_t, \tag{21}$$

where $\boldsymbol{g} = \nabla \mathcal{L}(\boldsymbol{\theta}_t)$ is the empirical gradient and $\boldsymbol{\xi}_t$ is a zero-mean stochastic noise term capturing minibatch sampling and finite-difference randomness.

To relate Eq. 21 to the true C-Flat gradient in Eq. 18, we decompose the deterministic surrogates as

$$A_t(\boldsymbol{g}_{vs,t-1}) = \boldsymbol{u}_t^{(s)} + \boldsymbol{\delta}_t^{(s)}, \tag{22}$$

$$B_t(\boldsymbol{g}_{vf,t-1}) = \boldsymbol{u}_t^{(f)} + \boldsymbol{\delta}_t^{(f)}, \tag{23}$$

where $\boldsymbol{u}_t^{(s)} := \boldsymbol{g}_s - \boldsymbol{g}$ and $\boldsymbol{u}_t^{(f)} := \boldsymbol{g}_f$, and $\boldsymbol{\delta}_t^{(s)}$, $\boldsymbol{\delta}_t^{(f)}$ are deterministic approximation errors (given $\mathcal{F}_t$) coming from EMA tracking, orthogonal decomposition residuals, and finite differences.

Substituting into Eq. 21 and comparing with Eq. 18, we obtain

$$\begin{aligned}\tilde{\boldsymbol{g}}_t &= \boldsymbol{g} + \big(\boldsymbol{u}_t^{(s)} + \boldsymbol{\delta}_t^{(s)}\big) + \lambda\big(\boldsymbol{u}_t^{(f)} + \boldsymbol{\delta}_t^{(f)}\big) + \boldsymbol{\xi}_t \\ &= \underbrace{\big(\boldsymbol{g} + \boldsymbol{u}_t^{(s)} + \lambda \boldsymbol{u}_t^{(f)}\big)}_{=\boldsymbol{G}_t} + \underbrace{\big(\boldsymbol{\delta}_t^{(s)} + \lambda \boldsymbol{\delta}_t^{(f)}\big)}_{:=\boldsymbol{b}_t} + \boldsymbol{\xi}_t \\ &= \boldsymbol{G}_t + \boldsymbol{b}_t + \boldsymbol{\xi}_t.\end{aligned} \tag{24}$$

Assumption 1 (Generalized smoothness)**.** *There exists $L > 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$ and any generalized gradients $\boldsymbol{g} \in \partial\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta})$, $\boldsymbol{g}' \in \partial\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}')$,*

$$\|\boldsymbol{g} - \boldsymbol{g}'\| \le L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

Assumption 2 (Bounded gradients and noise)**.** *There exist constants $G_{\max}, \Sigma \ge 0$ such that for all t and any $\boldsymbol{g} \in \partial\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t)$,*

$$\|\boldsymbol{g}\| \le G_{\max}, \qquad \mathbb{E}[\boldsymbol{\xi}_t \mid \mathcal{F}_t] = \boldsymbol{0}, \qquad \mathbb{E}[\|\boldsymbol{\xi}_t\|^2 \mid \mathcal{F}_t] \le \Sigma^2.$$

Assumption 3 (Deterministic approximation bias)**.** *The bias $\boldsymbol{b}_t$ satisfies $\|\boldsymbol{b}_t\| \le \Delta_t$ almost surely, where $\Delta_t \ge 0$ is $\mathcal{F}_t$-measurable. In Turbo, Δ_t aggregates:*

1. *EMA tracking error in estimating $\boldsymbol{g}_s - \boldsymbol{g}$ and $\boldsymbol{g}_f$ via $\boldsymbol{g}_{vs,t}$ and $\boldsymbol{g}_{vf,t}$;*
2. *orthogonal decomposition residuals in constructing the direction-invariant sharpness and flatness component $\boldsymbol{g}_{vs}$ and $\boldsymbol{g}_{vf}$;*
3. *finite-difference approximation errors in the proxy and proxy-perturbed gradients $\boldsymbol{g}_0, \boldsymbol{g}_1$.*

Assumption 4 (Stepsize and perturbation schedule)**.** *We use the schedule in [61]:*

$$\eta_t = \frac{\eta_0}{\sqrt{t}}, \qquad \rho_t = \frac{\rho_0}{\sqrt{t}},$$

with η_0 chosen so that $\eta_t \le 1/(4L)$ for all t.

Finally, the EMA tracking of $\boldsymbol{u}_t^{(s)}$ and $\boldsymbol{u}_t^{(f)}$ yields a bias sequence Δ_t satisfying

$$\sum_{t=1}^{T} \eta_t \Delta_t^2 \;\le\; C_{\Delta,1}\sqrt{T} + C_{\Delta,2}\log T \tag{25}$$

for some constants $C_{\Delta,1}, C_{\Delta,2} \ge 0$.

Justification of the deterministic bias assumption. The weighted bound on Δ_t in Eq. 25 is supported by the empirical directional stability of the cached components. As illustrated in Figure 3 and 7 in the main text, the direction invariant components $\boldsymbol{g}_{vs,t}$ and $\boldsymbol{g}_{vf,t}$ exhibit substantially higher cosine similarity across consecutive iterations than the empirical gradient $\boldsymbol{g}_t$ and the proxy gradient $\boldsymbol{g}_{0,t}$. This high stability suggests that the surrogate approximations $A_t(\cdot)$ and $B_t(\cdot)$ remain accurate within Turbo intervals (i.e., $\|\boldsymbol{\delta}_t^{(s)}\|$ and $\|\boldsymbol{\delta}_t^{(f)}\|$ stay small), which in turn justifies modeling the resulting approximation bias Δ_t controlled.

Lemma 1 (Generalized descent lemma)**.** *Under the generalized smoothness assumption, for any update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t\tilde{\boldsymbol{g}}_t$ and any $\boldsymbol{G}_t \in \partial\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t)$,*

$$\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_{t+1}) \le \mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t) - \eta_t\langle \boldsymbol{G}_t, \tilde{\boldsymbol{g}}_t\rangle + \tfrac{L}{2}\eta_t^2\|\tilde{\boldsymbol{g}}_t\|^2.$$

Proof. By generalized smoothness and the descent inequality for L-smooth functions, we have

$$\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_{t+1}) \le \mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t) + \langle \boldsymbol{G}_t, \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\rangle + \tfrac{L}{2}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2.$$

Substitute $\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = -\eta_t\tilde{\boldsymbol{g}}_t$ and rearrange. □

For EMA tracking, the bias sequence induced by $\boldsymbol{g}_{vs,t}$ and $\boldsymbol{g}_{vf,t}$ is absorbed into Δ_t, and we simply assume that it satisfies the bound in Eq. 25, which is standard for exponential moving averages under mild regularity conditions.

Single-step expected decrease. Starting from the descent lemma and using $\tilde{\boldsymbol{g}}_t = \boldsymbol{G}_t + \boldsymbol{b}_t + \boldsymbol{\xi}_t$, we obtain

$$\begin{aligned}\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_{t+1}) \le \mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t) &- \eta_t\langle \boldsymbol{G}_t, \boldsymbol{G}_t + \boldsymbol{b}_t + \boldsymbol{\xi}_t\rangle \\ &+ \tfrac{L}{2}\eta_t^2\|\boldsymbol{G}_t + \boldsymbol{b}_t + \boldsymbol{\xi}_t\|^2.\end{aligned} \tag{26}$$

Taking full expectation, and using $\mathbb{E}[\langle \boldsymbol{G}_t, \boldsymbol{\xi}_t\rangle \mid \mathcal{F}_t] = 0$, we get

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_{t+1})] \le \mathbb{E}[\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t)] &- \eta_t\mathbb{E}[\|\boldsymbol{G}_t\|^2] \\ &- \eta_t\mathbb{E}[\langle \boldsymbol{G}_t, \boldsymbol{b}_t\rangle] + \tfrac{L}{2}\eta_t^2\mathbb{E}[\|\boldsymbol{G}_t + \boldsymbol{b}_t + \boldsymbol{\xi}_t\|^2].\end{aligned} \tag{27}$$

a) Bias inner product term: By Cauchy–Schwarz and $\|\boldsymbol{b}_t\| \le \Delta_t$, we have

$$|\langle \boldsymbol{G}_t, \boldsymbol{b}_t\rangle| \le \|\boldsymbol{G}_t\|\,\|\boldsymbol{b}_t\| \le \tfrac{1}{2}\|\boldsymbol{G}_t\|^2 + \tfrac{1}{2}\Delta_t^2.$$

Using $|\mathbb{E}[Z]| \le \mathbb{E}[|Z|]$, this implies

$$-\eta_t\mathbb{E}[\langle \boldsymbol{G}_t, \boldsymbol{b}_t\rangle] \le \tfrac{\eta_t}{2}\mathbb{E}\|\boldsymbol{G}_t\|^2 + \tfrac{\eta_t}{2}\Delta_t^2. \tag{28}$$

b) Second moment term: For any three vectors x, y, z, Jensen's inequality gives

$$\|x + y + z\|^2 \le 3\big(\|x\|^2 + \|y\|^2 + \|z\|^2\big).$$

Applying this with $x = \boldsymbol{G}_t$, $y = \boldsymbol{b}_t$ and $z = \boldsymbol{\xi}_t$, we obtain

$$\|\boldsymbol{G}_t + \boldsymbol{b}_t + \boldsymbol{\xi}_t\|^2 \le 3\|\boldsymbol{G}_t\|^2 + 3\|\boldsymbol{b}_t\|^2 + 3\|\boldsymbol{\xi}_t\|^2,$$

hence, using $\|\boldsymbol{b}_t\| \le \Delta_t$ and $\mathbb{E}[\|\boldsymbol{\xi}_t\|^2 \mid \mathcal{F}_t] \le \Sigma^2$,

$$\mathbb{E}[\|\boldsymbol{G}_t + \boldsymbol{b}_t + \boldsymbol{\xi}_t\|^2] \le 3\,\mathbb{E}\|\boldsymbol{G}_t\|^2 + 3\Delta_t^2 + 3\Sigma^2. \tag{29}$$

Substituting Eq. 28 and Eq. 29 into Eq. 27, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_{t+1})] &\le \mathbb{E}[\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t)] - \eta_t\mathbb{E}\|\boldsymbol{G}_t\|^2 + \tfrac{\eta_t}{2}\mathbb{E}\|\boldsymbol{G}_t\|^2 \\ &\quad + \tfrac{\eta_t}{2}\Delta_t^2 + \tfrac{L}{2}\eta_t^2\big(3\mathbb{E}\|\boldsymbol{G}_t\|^2 + 3\Delta_t^2 + 3\Sigma^2\big) \\ &= \mathbb{E}[\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t)] - \Big(\tfrac{\eta_t}{2} - \tfrac{3L}{2}\eta_t^2\Big)\mathbb{E}\|\boldsymbol{G}_t\|^2 \\ &\quad + \Big(\tfrac{\eta_t}{2} + \tfrac{3L}{2}\eta_t^2\Big)\Delta_t^2 + \tfrac{3L}{2}\eta_t^2\Sigma^2. \end{aligned}\tag{30}$$

Under the stepsize condition $\eta_t \le 1/(4L)$, we have

$$\tfrac{\eta_t}{2} - \tfrac{3L}{2}\eta_t^2 \;\ge\; \tfrac{\eta_t}{8},$$

and hence

$$\frac{\eta_t}{8}\,\mathbb{E}\|\boldsymbol{G}_t\|^2 \le \mathbb{E}[\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t)] - \mathbb{E}[\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_{t+1})] + \Psi_t,$$

where

$$\Psi_t := \Big(\tfrac{\eta_t}{2} + \tfrac{3L}{2}\eta_t^2\Big)\Delta_t^2 + \tfrac{3L}{2}\eta_t^2\Sigma^2.$$

Summation. Summing from $t = 1$ to T, we obtain

$$\sum_{t=1}^{T}\frac{\eta_t}{8}\mathbb{E}\|\boldsymbol{G}_t\|^2 \le \mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_1) - L^* + \sum_{t=1}^{T}\Psi_t,$$

where $L^* = \inf_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta})$. Hence

$$\sum_{t=1}^{T}\eta_t\mathbb{E}\|\boldsymbol{G}_t\|^2 \le 8\big(\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_1) - L^*\big) + 8\sum_{t=1}^{T}\Psi_t. \tag{31}$$

Using $\eta_t = \eta_0/\sqrt{t}$ and the assumption $\sum_{t=1}^{T}\eta_t\Delta_t^2 \le C_{\Delta,1}\sqrt{T} + C_{\Delta,2}\log T$, together with $\sum_{t=1}^{T}\eta_t^2 \le \eta_0^2(1+\log T)$, we obtain

$$\sum_{t=1}^{T}\Psi_t \le C_1\sqrt{T} + C_2\log T + C_3$$

for some constants C_1, C_2, C_3. Thus there exist C_A, C_B, C_C such that

$$\sum_{t=1}^{T}\eta_t\mathbb{E}\|\boldsymbol{G}_t\|^2 \le C_A + C_B\sqrt{T} + C_C\log T.$$

Since $\eta_T = \eta_0/\sqrt{T}$ and η_t is nonincreasing, we have

$$\eta_T\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{G}_t\|^2 \le \sum_{t=1}^{T}\eta_t\mathbb{E}\|\boldsymbol{G}_t\|^2,$$

so

$$\frac{\eta_0}{\sqrt{T}}\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{G}_t\|^2 \le C_A + C_B\sqrt{T} + C_C\log T.$$

Dividing by T yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{G}_t\|^2 \le \frac{C_A}{\eta_0\sqrt{T}} + \frac{C_B}{\eta_0} + \frac{C_C\log T}{\eta_0\sqrt{T}}.$$

Theorem 1 (Turbo convergence in surrogate steps)**.** *Under the above assumptions and the EMA tracking condition Eq. 25, there exist constants $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3 \ge 0$ such that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t)\|^2\big] \le \frac{\tilde{C}_1 + \tilde{C}_2\log T}{\sqrt{T}} + \tilde{C}_3.$$

Moreover, if the bias sequence satisfies $\Delta_t = O(1/\sqrt{t})$ so that $\sum_{t=1}^{T}\eta_t\Delta_t^2 = O(1+\log T)$, then $\tilde{C}_3 = 0$ and Turbo recovers a GAM-style rate [6]:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla\mathcal{L}_{\mathrm{CFlat}}(\boldsymbol{\theta}_t)\|^2\big] = O\Big(\frac{1+\log T}{\sqrt{T}}\Big).$$

A.10. C-Flat Turbo Algorithm

Algorithm 1 C-Flat Turbo

Input: Training phase T, training data S^T, model parameter $\boldsymbol{\theta}$, total iterations J, oracle loss function $\mathcal{L}$, learning rate η, C-Flat coefficient λ, Turbo step k, $\mu_{s,0} = \mu_{f,0} = 0$, $\sigma_{s,0} = \sigma_{f,0} = 10^{-8}$.
Output: Model trained at the current time T with Turbo.

for $j = 1$ to J, sample batch B_j^T from dataset S^T **do**
  Compute empirical gradient: $\boldsymbol{g} = \nabla\mathcal{L}(\boldsymbol{\theta})$
  Initialize update direction: $\overline{\boldsymbol{g}} = \boldsymbol{g}$
  Update EMA statistics: $\mu_{s,j}, \sigma_{s,j}, \mu_{f,j}, \sigma_{f,j}$ by Eq. 8
  if $\|\boldsymbol{g}\|^2 \ge \mu_{s,j} + \sigma_{s,j}$ **then**
    if $j \bmod k = 0$ **then**
      Compute SAM gradient $\boldsymbol{g}_s = \nabla\mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}_0^*)$ by Eq. 4
      Cache direction-invariant sharpness component $\boldsymbol{g}_{vs}$ by Eq. 6
    else
      Simulate sharpness increment using cached $\boldsymbol{g}_{vs}$, that means $\boldsymbol{g}_s = \boldsymbol{g} + \beta\frac{\|\boldsymbol{g}\|}{\|\boldsymbol{g}_{vs}\|}\boldsymbol{g}_{vs}$
    end if
    Update direction: $\overline{\boldsymbol{g}} = \boldsymbol{g}_s$
  end if
  Compute proxy model parameter $\boldsymbol{\theta}_p = \boldsymbol{\theta} + \boldsymbol{\epsilon}_1^*$ by Eq. 5
  Compute proxy gradient: $\boldsymbol{g}_0 = \nabla\mathcal{L}(\boldsymbol{\theta}_p)$
  if $\|\boldsymbol{g}_0\|^2 \ge \mu_{f,j} + \sigma_{f,j}$ **then**
    if $j \bmod k = 0$ **then**
      Compute proxy-perturbed gradient $\boldsymbol{g}_1 = \nabla\mathcal{L}\Big(\boldsymbol{\theta}_p + \rho'\frac{\boldsymbol{g}_0}{\|\boldsymbol{g}_0\|}\Big)$
      Compute first-order flatness gradient $\boldsymbol{g}_f$ from $\boldsymbol{g}_0, \boldsymbol{g}_1$ by Eq. 5
      Cache direction-invariant flatness component $\boldsymbol{g}_{vf}$ by Eq. 7
    else
      Simulate flatness direction using cached $\boldsymbol{g}_{vf}$, that means $\boldsymbol{g}_f = \boldsymbol{g}_0 + \beta\frac{\|\boldsymbol{g}_0\|}{\|\boldsymbol{g}_{vf}\|}\boldsymbol{g}_{vf}$
    end if
    Update direction: $\overline{\boldsymbol{g}} = \overline{\boldsymbol{g}} + \lambda\cdot\boldsymbol{g}_f$
  end if
  Optionally update the Turbo step k according to Section 4.6
  Update model parameter: $\boldsymbol{\theta}^T = \boldsymbol{\theta}^T - \eta\cdot\overline{\boldsymbol{g}}$
end for