A PROOF OF THEOREM 1

Theorem. *Let $\mathcal{L}^{\text{aux}}(\theta_t)$ and $\mathcal{L}^{\text{prim}}(\theta_t)$ represent the full batch losses of the auxiliary tasks and primary task respectively at step t. We assume the gradients of $\mathcal{L}^{\text{aux}}$ and $\mathcal{L}^{\text{prim}}$ are Lipschitz continuous with constant $L > 0$. Following the update rule : $\theta_{t+1} = \theta_t - \alpha \cdot \tilde{\boldsymbol{g}}_{aux}$, where $\alpha \leq \frac{1}{L}$ is the learning rate, we are guaranteed :*

$$\mathcal{L}^{\text{aux}}(\theta_{t+1}) \leq \mathcal{L}^{\text{aux}}(\theta_t)$$
$$\mathcal{L}^{\text{prim}}(\theta_{t+1}) \leq \mathcal{L}^{\text{prim}}(\theta_t)$$

If $\eta_- = 0$ and $\eta_\perp, \eta_+ \geq 0$

Proof. Let $\boldsymbol{V}_t \in \mathbb{R}^{k \times D}$ be the orthonormal matrix whose rows span the per-example primary task gradients $\boldsymbol{J}^*$ at timestep t. The projections of the average primary task gradient $\boldsymbol{g}_{prim} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{J}^*_{i,:}$ and average auxiliary task gradient $\boldsymbol{g}_{aux}$ at iteration t are

$$\boldsymbol{p}_{prim} = \boldsymbol{V}_t\big(\boldsymbol{g}_{prim}\big)^T$$
$$\boldsymbol{p}_{aux} = \boldsymbol{V}_t\big(\boldsymbol{g}_{aux}\big)^T$$

$\boldsymbol{p}_{prim}$ and $\boldsymbol{p}_{aux}$ will agree on some directions (same sign on those components). We use the operator $[x]_+$ to mark these directions of agreement. This operator preserves components that agree and sets those that disagree to zero. As an example given $\boldsymbol{p}_{prim} = [1, 1, -1]$ and $\boldsymbol{p}_{aux} = [1, 3, 10]$, $[\boldsymbol{p}_{prim}]_+ = [1, 1, 0]$ and $[\boldsymbol{p}_{aux}]_+ = [1, 3, 0]$. For directions that disagree (different signs of the respective components), we introduce the operator $[x]_-$. In the above example $[\boldsymbol{p}_{prim}]_- = [0, 0, -1]$ and $[\boldsymbol{p}_{aux}]_- = [0, 0, 10]$. Note that our operators are defined by comparing two vectors x_1 and x_2, Our operators have the following properties by definition :

$$x = [x]_- + [x]_+$$

and

$$[x]_+ \perp [x]_-, [x_1]_\pm \perp [x_2]_\mp$$

From Equation 2 :

$$\tilde{\boldsymbol{g}}_{aux} = \eta_+ \boldsymbol{g}^+_{aux} + \eta_- \boldsymbol{g}^-_{aux} + \eta_\perp \boldsymbol{g}^\perp_{aux}$$

We can re-write this in terms of $[x]_\pm$ as :

$$\tilde{\boldsymbol{g}}_{aux} = \eta_+[\boldsymbol{p}_{aux}]_+ + \eta_-[\boldsymbol{p}_{aux}]_- + \eta_\perp\big(\boldsymbol{g}_{aux} - \boldsymbol{p}_{aux}\big)$$

We now proceed to show the effect of the gradient descent update below on $\mathcal{L}^{\text{aux}}(\theta_{t+1})$ and $\mathcal{L}^{\text{prim}}(\theta_{t+1})$.

$$\theta_{t+1} = \theta_t - \alpha \cdot \tilde{\boldsymbol{g}}_{aux} \tag{3}$$

How does this update affect the loss on the primary task loss $\mathcal{L}^{\text{prim}}(\theta_{t+1})$?

$\mathcal{L}^{\text{prim}}(\theta_{t+1}) = \mathcal{L}^{\text{aux}}(\theta_t - \alpha \cdot \tilde{\boldsymbol{g}}_{aux})$

$$\begin{aligned}
&\approx \mathcal{L}^{\text{prim}}(\theta_t) - \alpha\big(\tilde{\boldsymbol{g}}_{aux}\big)^T \boldsymbol{g}_{prim} \;\; \textit{(First order Taylor Expansion)} \\
&= \mathcal{L}^{\text{prim}}(\theta_t) - \alpha\bigg(\eta_+[\boldsymbol{p}_{aux}]_+ + \eta_-[\boldsymbol{p}_{aux}]_- + \eta_\perp \boldsymbol{g}^\perp_{aux}\bigg)^T \boldsymbol{g}_{prim} \\
&= \mathcal{L}^{\text{prim}}(\theta_t) - \alpha\bigg(\eta_+[\boldsymbol{p}_{aux}]_+ + \eta_-[\boldsymbol{p}_{aux}]_- + \eta_\perp \boldsymbol{g}^\perp_{aux}\bigg)^T \bigg([\boldsymbol{p}_{prim}]_+ + [\boldsymbol{p}_{prim}]_-\bigg) \\
&= \mathcal{L}^{\text{prim}}(\theta_t) - \alpha\bigg(\eta_+\bigg([\boldsymbol{p}_{aux}]_+^T[\boldsymbol{p}_{prim}]_+ + [\boldsymbol{p}_{aux}]_+^T[\boldsymbol{p}_{prim}]_-\bigg) + \eta_-\bigg([\boldsymbol{p}_{aux}]_-^T[\boldsymbol{p}_{prim}]_+ + [\boldsymbol{p}_{aux}]_-^T[\boldsymbol{p}_{prim}]_-\bigg)\bigg) \\
&= \mathcal{L}^{\text{prim}}(\theta_t) - \alpha\bigg(\eta_+[\boldsymbol{p}_{aux}]_+^T[\boldsymbol{p}_{prim}]_+ + \eta_-[\boldsymbol{p}_{aux}]_-^T[\boldsymbol{p}_{prim}]_-\bigg) \\
&\leq \mathcal{L}^{\text{prim}}(\theta_t) \;\; \textit{(if } \eta_- \leq 0,\ \eta_\perp, \eta_+ \geq 0\textit{)}
\end{aligned}$$

Note that in going from line 3 to 4 in the proof above, we use the fact that $\left(\boldsymbol{g}^{\perp}_{aux}\right)^T \boldsymbol{g}_{prim} = 0$ since $\boldsymbol{g}^{\perp}_{aux}$ lies outside the subspace and $\boldsymbol{g}_{prim}$ lies inside it. For the last step of the proof, we use the observations below :

$$
\begin{aligned}
[\boldsymbol{p}_{aux}]_{+}[\boldsymbol{p}_{prim}]_{+} &\geq 0 \text{ since these directions agree in sign} \\
[\boldsymbol{p}_{aux}]_{-}[\boldsymbol{p}_{prim}]_{-} &\leq 0 \text{ since these directions disagree in sign} \\
[\boldsymbol{p}_{aux}]_{+}[\boldsymbol{p}_{prim}]_{-} &= 0 \text{ by the property of the } [x]_{\pm} \text{ operator} \\
[\boldsymbol{p}_{aux}]_{-}[\boldsymbol{p}_{prim}]_{+} &= 0 \text{ same motivation as above}
\end{aligned}
$$

How does Equation 3 affect the auxiliary task loss $\mathcal{L}^{\text{aux}}(\theta_{t+1})$?

$$
\begin{aligned}
\mathcal{L}^{\text{aux}}(\theta_{t+1}) &= \mathcal{L}^{\text{aux}}(\theta_t - \alpha \cdot \tilde{\boldsymbol{g}}_{aux}) \\
&\approx \mathcal{L}^{\text{aux}}(\theta_t) - \alpha\big(\tilde{\boldsymbol{g}}_{aux}\big)^T \boldsymbol{g}_{aux} \;\; \textit{(First order Taylor Expansion)} \\
&= \mathcal{L}^{\text{aux}}(\theta_t) - \alpha\big(\eta_{\perp}\boldsymbol{g}^{\perp}_{aux} + \eta_{+}\boldsymbol{g}^{+}_{aux} + \eta_{-}\boldsymbol{g}^{-}_{aux}\big)^T\big(\boldsymbol{g}^{\perp}_{aux} + \boldsymbol{g}^{+}_{aux} + \boldsymbol{g}^{-}_{aux}\big) \\
&= \mathcal{L}^{\text{aux}}(\theta_t) - \alpha\big(\eta_{\perp}\|\boldsymbol{g}^{\perp}_{aux}\|^2 + \eta_{+}\|\boldsymbol{g}^{+}_{aux}\|^2 + \eta_{-}\|\boldsymbol{g}^{-}_{aux}\|^2\big) \;\; \textit{(Cross terms cancel due to orthogonality)} \\
&\leq \mathcal{L}^{\text{aux}}(\theta_t) \;\; \textit{(If } \eta_{-}, \eta_{\perp}, \eta_{+} \geq 0\textit{)}
\end{aligned}
$$

Thus, choosing $\eta_{-} = 0$ ensures that we are minimizing both $\mathcal{L}^{\text{aux}}(\theta_t)$ and $\mathcal{L}^{\text{prim}}(\theta_t)$. We can combine this with the constraint on $\alpha \leq \frac{1}{L}$ to derive convergence guarantees after some T steps as in optimization literature. □

B RANDOMIZED MATRIX THEORY

```
Algorithm 2: randomized_lowrank_approx : Construct low rank approximation
Require : $\boldsymbol{J} \in \mathbb{R}^{m \times D}$ : Input Matrix
Require : $k$ : Rank of subspace
    $\Pi \sim \mathcal{N}(0, I) \in \mathbb{R}^{k \times m}$
    $\boldsymbol{C} = \Pi \boldsymbol{J}$
    $\boldsymbol{V} \leftarrow \text{Gram\_Schmidt}(\text{C})$
Return : $\boldsymbol{V} \in \mathbb{R}^{k \times D}$ : Low rank approximation of $\boldsymbol{J}$
```

The Gram_Schmidt procedure orthogonalizes the rows of an input matrix.

C MORE EXPERIMENTAL DETAILS

Image Classification For MultiCifar100, unlike Rosenbaum et al. (2017); Yu et al. (2020) who use a 500-100 train-test split for examples under each fine-grained CIFAR 100 label, we include a validation set and therefore opt for a 400-100-100 train-validation-test split. We test on all 1000 test examples per class.

For Cat-vs-Dog, we use 100 examples from the training set as validation and test on all 1000 test examples per-class.

For Image Classification experiments, we perform pre-training with a learning rate of 1e-4 for all experiments and finetuning learning rate of 5e-4. These values were selected after coarse hyper-parameter search. In both pre-training and finetuning settings, we decay the learning rate by 0.5 if the validation loss has not improved over 4 epochs, up till a minimum learning rate of 1e-5. we use the Adam Optimizer (Kingma & Ba, 2014) with $\beta = (0.9, 0.999)$. We clip all gradient norms to 1.0 before performing gradient descent. We cross-validated dropout rates within the set {0.05, 0.1, 0.2, 0.3} for both pre-training and finetuning steps. We cross validate η_{prim} based on the relative sizes of primary and auxilary task datasets. All experiments are averaged over 5 random seeds.

Medical Imaging Transfer Table 5 presents a more detailed breakdown of the ChexPert task. For 0.5% of Imagenet, our best performing configuration was $\eta_{aux} = (1.0, 0.0, -1.0)$. We did not use the primary task gradient directly for pre-training so $\eta_{prim} = 0.0$ for all cases. For ATTITUD, we use the same learning rates as in the Image classification setup above. For the No-Pretraining and Vanilla pretraining we cross-validated the learning rates for both finetuning and pre-training from the set {1e-3, 1e-4}. We cross-validated the same list of dropout values above.

Method	No-Pretraining	Pretrained w Imgnet	Ours (0.5% Imgnet)	Ours (1% Imgnet)
Atelectasis	76.0 ± 1.82	79.0 ± 3.66	81.6 ± 1.38	**81.8±0.80**
Cardiomegaly	74.9 ± 2.34	75.8 ± 4.04	78.0 ± 2.13	**80.7±1.79**
Consolidation	83.2 ± 2.26	**85.3±1.86**	**85.6±2.32**	84.9 ± 1.36
Edema	79.5 ± 1.27	82.6 ± 0.76	**85.2±1.23**	84.7 ± 1.78
P. Effusion	77.9 ± 1.88	**84.4±0.75**	83.4 ± 1.80	**84.3±0.65**

Table 5: Results on ChexPert-5k tasks measured by average AUC (Area Under Roc-Curve)

Text Classification For our NLP experiments, we tried limiting the number of layers we applied ATTITUD to. We achieved good performance without applying ATTITUD to the word embedding layers (these were updated with untouched auxiliary task gradients). We cross-validated $\eta_{prim} = \{0.01, 0.05, 0.0025\}$

For all experiments involving ATTITUD, We cross-validate the following choices of the subspace size $k \in \{5, 10, 20\}$ from $\boldsymbol{J}^* \in \mathbb{R}^{m \times D}$ using $m \in \{32, 64\}$. We recompute the subspace every 10 steps for vision experiments and every 4 steps for NLP experiments. We performed early stopping for all experiments if no improvement after 10 epochs.

Ablation of Fraction of Norm within Subspace The left pane of Figure 3 reinforces our intuition and confirms that our choice of the top-k singular vectors (*randomized_svd*) gives the best accuracy as averaged across 5 seeds. *random* is the basis spanned by k randomly chosen orthogonal vectors in $\mathbb{R}^D$, *unit_avg_grad* is the basis spanned by the average primary task gradient whilst *canonical* uses the per-parameter basis. We use the fraction of the norm of sample gradients within a subspace as indicators of how *semantically* meaningful that choice of subspace is. We expect that a *semantically* meaningful choice of basis will achieve better generalization performance because it captures the essential parts of the gradient with $k \ll D$. *canonical* trivially captures all the norm of the sampled gradient vectors but because $k = D$, it generalizes poorly. Notice that only small fractions of the norms of sample primary and auxiliary task average gradients lie in the subspace for *random* and *unit_avg_grad*, whilst significant fractions lie in *randomized_svd*.

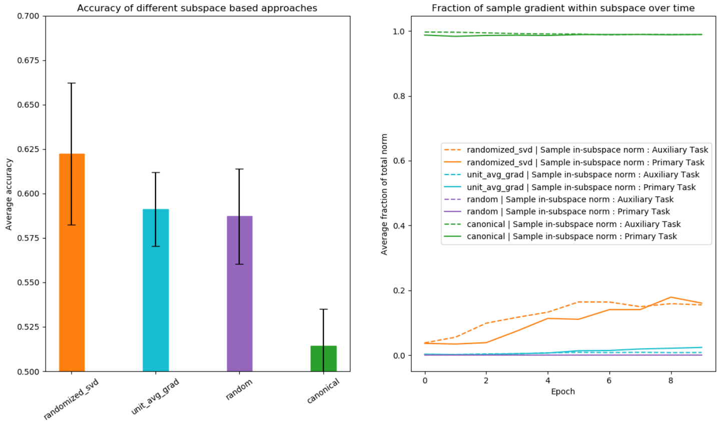

Figure 3: Experiment conducted on Cat-vr-Dog Cifar10 dataset. We use subspace size $k = 5$. **Left** Averaged accuracy across 5 seeds of different choices of basis. Our choice, randomized_svd performs best. **Right** We look at the fraction of the norm of $\boldsymbol{g}_{aux}$ within each subspace (dashed line). We also do so for a randomly sampled mini-batch of the primary task (solid line).