Independent Component Alignment for Multi-Task Learning Supplemental Material

Dmitry Senushkin Nikolay Patakin Arseny Kuznetsov Anton Konushin
Samsung Research
{d.senushkin, n.patakin, a.konushin}@samsung.com

A. Convergence Analysis

Synopsis. In these theorems, we prove that the worst case performance of *Aligned-MTL* and *Aligned-MTL-U*B approaches is no worse than of standard gradient descent. The constraints mentioned in convergence theorems below are mild enough to be satisfied in practice. Our approach converges to a Pareto-stationary point with pre-defined tasks weights, thus providing more control over an optimization result.

Lemma 1 *Assume $\mathcal{L}(\boldsymbol{\theta})$ to be continuously differentiable and $\nabla\mathcal{L}(\boldsymbol{\theta})$ to be Lipschitz continuous with $\Lambda > 0$. Then, the following restriction holds for a gradient descent with a step size α and an update rule $\boldsymbol{r}$:*

$$\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) \geq \alpha\langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{r} \rangle - \frac{\alpha^2\Lambda}{2}\|\boldsymbol{r}\|^2. \quad (1)$$

Proof *Let us consider a gradient descent $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\delta}$, where $\boldsymbol{\delta} = -\alpha\boldsymbol{r}$. From the fundamental theorem of calculus, we derive:*

$$\mathcal{L}(\boldsymbol{\theta}_t + \boldsymbol{\delta}) - \mathcal{L}(\boldsymbol{\theta}_t) = \int_0^1 \langle \nabla\mathcal{L}(\boldsymbol{\theta}_t + s\boldsymbol{\delta}), \boldsymbol{\delta} \rangle \,\mathrm{d}s. \quad (2)$$

By adding and subtracting the value $\langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \delta \rangle = \int_0^1 \langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \delta \rangle \,\mathrm{d}s$, we obtain:

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) - \mathcal{L}(\boldsymbol{\theta}_t) = \langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{\delta} \rangle + \quad (3)$$

$$+ \int_0^1 \langle \nabla\mathcal{L}(\boldsymbol{\theta}_t + s\boldsymbol{\delta}) - \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{\delta} \rangle \,\mathrm{d}s. \quad (4)$$

Since the gradient satisfies the Lipschitz condition $\|\nabla\mathcal{L}(\boldsymbol{\theta}_t + s\boldsymbol{\delta}) - \nabla\mathcal{L}(\boldsymbol{\theta}_t)\| \leq \Lambda\|\boldsymbol{\theta}_t + s\boldsymbol{\delta} - \boldsymbol{\theta}_t\|$ and due to inequality $\langle x, y \rangle \leq \|x\|\|y\|$, we can transform the integral as following:

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) - \mathcal{L}(\boldsymbol{\theta}_t) = \langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{\delta} \rangle +$$
$$+ \int_0^1 \langle \nabla\mathcal{L}(\boldsymbol{\theta}_t + s\boldsymbol{\delta}) - \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{\delta} \rangle \,\mathrm{d}s \leq$$
$$\langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{\delta} \rangle + \int_0^1 \Lambda\|\boldsymbol{\theta}_t + s\boldsymbol{\delta} - \boldsymbol{\theta}_t\|\|\boldsymbol{\delta}\|\mathrm{d}s \leq$$
$$-\alpha\langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{r} \rangle + \Lambda\int_0^1 \|-s\alpha\boldsymbol{r}\|_2\| - \alpha\boldsymbol{r}\|_2\mathrm{d}s \leq$$
$$-\alpha\langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{r} \rangle + \alpha^2\Lambda\|\boldsymbol{r}\|^2\int_0^1 s\mathrm{d}s \leq$$
$$-\alpha\langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{r} \rangle + \alpha^2\Lambda\|\boldsymbol{r}\|^2$$

Therefore, we obtain the final constraint:

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) - \mathcal{L}(\boldsymbol{\theta}_t) \leq -\alpha\langle \nabla\mathcal{L}(\boldsymbol{\theta}_t), \boldsymbol{r} \rangle + \alpha^2\Lambda\|\boldsymbol{r}\|^2. \quad (5)$$

Theorem 1 (Aligned-MTL) *Assume $\mathcal{L}_0(\boldsymbol{\theta}), \ldots, \mathcal{L}_T(\boldsymbol{\theta})$ are lower-bounded continuously differentiable functions with Lipschitz continuous gradients with $\Lambda > 0$. A gradient descent with an aligned gradient and a step size $\alpha \leq \frac{1}{\Lambda}$ converges linearly to a Pareto-stationary point where $\nabla\mathcal{L}_0(\boldsymbol{\theta}) = 0$.*

Proof (Aligned-MTL) *Given the aforementioned assumptions, the cumulative objective satisfies Lemma 1 with $\boldsymbol{r} = \hat{\boldsymbol{G}}\boldsymbol{w} = \hat{g}_0$ and $\nabla\mathcal{L}_0(\boldsymbol{\theta}) = \boldsymbol{G}\boldsymbol{w} = \boldsymbol{g}_0$:*

$$\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) \geq \alpha\boldsymbol{g}_0^\top\hat{\boldsymbol{g}}_0 - \frac{\alpha^2\Lambda}{2}\|\hat{\boldsymbol{g}}_0\|^2. \quad (6)$$

According to SVD, $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1, \ldots, \sigma_R\}$ where $R = \mathrm{rank}\,\boldsymbol{G}$, and $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$. By definition of the Aligned-MTL, we get:

$$\boldsymbol{g}_0^\top\hat{\boldsymbol{g}}_0 = \sigma_R\boldsymbol{w}^\top\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{V}^\top\boldsymbol{w} =$$
$$= \sigma_R\boldsymbol{w}^\top\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{w} = \sum_{r=1}^{R}\sigma_R\sigma_r(\boldsymbol{w}^\top\boldsymbol{v}_r)^2$$

Similarly, $\|\hat{\boldsymbol{g}}_0\|^2 = \sum_{r=1}^{R} \sigma_R^2 (\boldsymbol{w}^\top \boldsymbol{v}_r)^2$. *Since* $\alpha \leq \frac{1}{\Lambda}$ *and* $\boldsymbol{w}^\top \boldsymbol{v}_r > \varepsilon$, *Eq.* (6) *can be further bounded:*

$$\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) \geq \sigma_R^2 \frac{\alpha}{2} \underbrace{\sum_{r=1}^{R} \underbrace{\left(2\frac{\sigma_r}{\sigma_R} - 1\right)}_{>1} \left(\boldsymbol{w}^\top \boldsymbol{v}_r\right)^2}_{>\|\boldsymbol{V}\boldsymbol{w}\|^2 > \varepsilon^2} > \frac{\alpha \sigma_R^2}{2} \frac{\varepsilon^2}{\sigma_1^2} \sigma_1^2.$$

The dominance is always finite: $\frac{\sigma_R}{\sigma_1} > C$. *Moreover,* $\sigma_1 = \max_{\boldsymbol{x} \neq 0} \frac{\|\boldsymbol{G}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$, *therefore* $\sigma_1 \geq \frac{\|\boldsymbol{g}_0\|}{\|\boldsymbol{w}\|}$. *Respectively:*

$$\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) > \frac{\alpha \varepsilon^2 C^2}{2\|\boldsymbol{w}\|^2} \|\boldsymbol{g}_0\|^2. \tag{7}$$

The sequence of $\mathcal{L}(\boldsymbol{\theta}_t)$ *is monotonically decreasing and bounded (under assumption), and hence converging. Then* $\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) \to 0$ *if* $t \to \infty$. *Thereby, we have a local convergence of the gradient descent:*

$$\|\boldsymbol{g}_0\|^2 < \frac{2\|\boldsymbol{w}\|^2}{\alpha C^2 \epsilon^2} \Big(\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1})\Big) \to 0 \quad as \quad t \to \infty. \tag{8}$$

The same estimate appears in case of the gradient descent. Accordingly, the convergence of Aligned-MTL is similar to that of the gradient descent, i.e. *linear –* $\mathcal{O}(\frac{1}{T})$.

Theorem 2 (A-MTL-UB) *Assume* $\mathcal{L}_0(\boldsymbol{\theta}), \ldots, \mathcal{L}_T(\boldsymbol{\theta})$ *are lower-bounded continuously differentiable functions with Lipschitz continuous gradients with* $\Lambda > 0$. *Suppose* $\boldsymbol{J} = \frac{\partial \boldsymbol{H}}{\partial \boldsymbol{\theta}}$ *to be a full rank,* i.e. $\operatorname{rank} \boldsymbol{J} = \min\{|\boldsymbol{\theta}|, |\boldsymbol{H}|\}$. *A gradient descent with an aligned gradient and a step size* $\alpha \leq \frac{1}{\Lambda}$ *converges linearly to a Pareto-stationary point where* $\nabla \mathcal{L}_0(\boldsymbol{\theta}) = 0$.

Proof (Aligned-MTL-UB) *Similarly to the Theorem 1, under the aforementioned assumptions, the cumulative objective satisfies Lemma 1 with* $\boldsymbol{r} = \sigma_R \boldsymbol{J}\hat{\boldsymbol{Z}}\boldsymbol{w} = \hat{\boldsymbol{g}}_0$ *and* $\nabla \mathcal{L}_0(\boldsymbol{\theta}) = \boldsymbol{J}\boldsymbol{Z}\boldsymbol{w} = \boldsymbol{g}_0$:

$$\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) \geq \alpha \boldsymbol{g}_0^\top \hat{\boldsymbol{g}}_0 - \frac{\alpha^2 \Lambda}{2} \|\hat{\boldsymbol{g}}_0\|^2. \tag{9}$$

According to SVD, $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, $\boldsymbol{\Sigma} = \operatorname{diag}\{\sigma_1, \ldots, \sigma_R\}$ *where* $R = \operatorname{rank} \boldsymbol{Z}$, *and* $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$. *By definition of the Aligned-MTL-UB, we get:*

$$\boldsymbol{g}_0^\top \hat{\boldsymbol{g}}_0 = \sigma_R \boldsymbol{w}^\top \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top \boldsymbol{J}^\top \boldsymbol{J}\boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{w}$$
$$\hat{\boldsymbol{g}}_0^\top \hat{\boldsymbol{g}}_0 = \sigma_R^2 \boldsymbol{w}^\top \boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{J}^\top \boldsymbol{J}\boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{w}$$

Since $\boldsymbol{J}$ *is full rank,* $\boldsymbol{J}^\top \boldsymbol{J}$ *is positive definite. Any positive definite matrix is congruent to a diagonal* ($\boldsymbol{D}$) *with positive and ordered eigenvalues on the main diagonal. Thus, replacing all eigenvalues* λ_i^2 *with the smallest one* λ_K^2 *does not increase the inner product produced by this matrix:* $\boldsymbol{x}\boldsymbol{D}\boldsymbol{x} \geq \lambda_K \boldsymbol{x}^\top \boldsymbol{x}$. *By taking this into consideration, we can bound the right side of Eq.* (9)*:*

$$\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) \geq \frac{\alpha}{2}(2\boldsymbol{g}_0 - \hat{\boldsymbol{g}}_0)^\top \hat{\boldsymbol{g}}_0 \geq$$
$$\frac{\alpha}{2}(2\sigma_R \boldsymbol{w}^\top \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top - \sigma_R^2 \boldsymbol{w}^\top \boldsymbol{V}\boldsymbol{U}^\top)\boldsymbol{J}^\top \boldsymbol{J}\boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{w} \geq$$
$$\sigma_R^2 \lambda_K^2 \underbrace{\sum_{r=1}^{R} \underbrace{\left(2\frac{\sigma_r}{\sigma_R} - 1\right)}_{>1} \left(\boldsymbol{w}^\top \boldsymbol{v}_r\right)^2}_{>\|\boldsymbol{V}\boldsymbol{w}\|^2 > \varepsilon^2}$$

Thus:

$$\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) \geq \frac{\alpha \varepsilon^2 \sigma_R^2 \lambda_K^2}{2\sigma_1^2 \lambda_1^2} \sigma_1^2 \lambda_1^2. \tag{10}$$

Following the assumption, $\frac{\sigma_R}{\sigma_1} > C_\sigma$ *and* $\frac{\lambda_K}{\lambda_1} > C_\lambda$. *Moreover,* $\sigma_1 = \max_{\boldsymbol{x} \neq 0} \frac{\|\boldsymbol{Z}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \geq \frac{\|\boldsymbol{Z}\boldsymbol{w}\|}{\|\boldsymbol{w}\|}$ *and* $\lambda_1 = \|\boldsymbol{J}\|$. *Therefore, we obtain the final bound:*

$$\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) \geq \frac{\alpha \varepsilon^2 C_\sigma^2 C_\lambda^2}{2\|\boldsymbol{w}\|^2} \|\boldsymbol{G}_{\boldsymbol{Z}}\boldsymbol{w}\|^2 \|\boldsymbol{J}\|^2 \geq$$
$$\frac{\alpha \varepsilon^2 C_\sigma^2 C_\lambda^2}{2\|\boldsymbol{w}\|^2} \|\boldsymbol{g}_0\|^2.$$

The sequence of $\mathcal{L}(\boldsymbol{\theta}_t)$ *is monotonically decreasing and bounded (under assumption), and hence converging. Then* $\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1}) \to 0$ *if* $t \to \infty$. *Thereby, we have a local convergence of the gradient descent:*

$$\|\boldsymbol{g}_0\|^2 < \frac{2\|\boldsymbol{w}\|^2}{\alpha C_\sigma^2 C_\lambda^2 \varepsilon^2} \Big(\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_{t+1})\Big) \to 0 \quad as \quad t \to \infty. \tag{11}$$

B. Condition Number

The stability criterion is closely related to the dominance and conflicts. We can find a functional dependence between them for some special cases: **a)** gradients $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ have equal magnitude but not orthogonal, **b)** they are othogonal but have different norms. To this end, we formulate the following colloraries.

Collorary 1 *Given* $\boldsymbol{g}_1 \perp \boldsymbol{g}_2$ *condition number* κ *is*

$$\kappa = \max\left\{\frac{\|\boldsymbol{g}_1\|}{\|\boldsymbol{g}_2\|}, \frac{\|\boldsymbol{g}_2\|}{\|\boldsymbol{g}_1\|}\right\}$$

Proof *By initial assumtions the Gram matrix* $\boldsymbol{G}^\top \boldsymbol{G}$ *is diagonal:*

$$\boldsymbol{G}^\top \boldsymbol{G} = \operatorname{diag}\{\|\boldsymbol{g}_1\|^2, \|\boldsymbol{g}_2\|^2\}$$

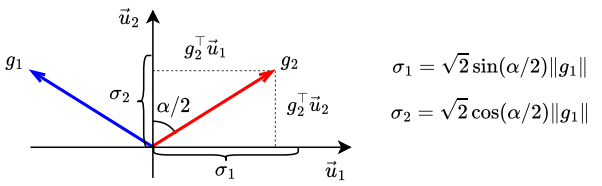


Figure 1. The condition number depends on the angle between gradient vectors. Due to the symmetry one of the principal components is a bisectrix of this angle.

At the same time, this matrix can be factorized using eigen decomposition:

$$\boldsymbol{G}^\top \boldsymbol{G} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top, \quad \boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I}, \quad \boldsymbol{\Sigma} = \text{diag}\{\sigma_1, \sigma_2\}$$

Thus, the singular values are proportional to the gradient magnitudes up to a symmetric swap to keep ordering of singular values. The coefficient of proportionality is not valuable, since the condition number is invariant to the global scale. Therefore, we derive:

$$\kappa = \max\left\{\frac{\|\boldsymbol{g}_1\|}{\|\boldsymbol{g}_2\|}, \frac{\|\boldsymbol{g}_2\|}{\|\boldsymbol{g}_1\|}\right\}$$

Collorary 2 *Given $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ with equal magnitudes,* i.e. *$\|\boldsymbol{g}_1\| = \|\boldsymbol{g}_2\|$, and with α angle in between the condition number κ is*

$$\kappa = \begin{cases} \tan(\alpha/2) & \frac{\pi}{4} < \alpha/2 \leq \frac{\pi}{2} \\ \text{ctan}(\alpha/2) & 0 < \alpha/2 < \frac{\pi}{4} \end{cases} \tag{12}$$

Proof *The direct collorary of SVD states, that the principal components $\boldsymbol{u}_i$ are direction with maximum norm of projections over all gradients. Formally:*

$$\sigma_1 = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{G}^\top \boldsymbol{x}\| = \|\boldsymbol{G}^\top \boldsymbol{u}_1\|$$
$$\sigma_2 = \max_{\|\boldsymbol{x}\|=1, \boldsymbol{x}\perp\boldsymbol{u}_1} \|\boldsymbol{G}^\top \boldsymbol{x}\| = \|\boldsymbol{G}^\top \boldsymbol{u}_2\|$$

Since the gradients have the same length, one of the principal components is the bisectrix of angle between them. For clarity, we suppose, that the bisectrix is the second component. Then, the singular values can be computed trivially (Fig. 1):

$$\sigma_1 = \sqrt{2}\sin(\alpha/2)\|\boldsymbol{g}_1\|$$
$$\sigma_2 = \sqrt{2}\cos(\alpha/2)\|\boldsymbol{g}_1\|$$

Accroding to these expressions the condition number is tangent or cotangent up to a symmetric swap to keep ordering of singular values. In orthoginal case, the condition number is unit.

C. Synthetic Example

The synthetic example is a two-task objective containing areas with the presence of conflicting and dominating gradients between loss components. Formally, we use the same objective as in previous works [2, 5]:

$$\mathcal{L}_1 = c_1(\boldsymbol{\theta})f_1(\boldsymbol{\theta}) + c_2(\boldsymbol{\theta})g_1(\boldsymbol{\theta})$$
$$\mathcal{L}_2 = c_1(\boldsymbol{\theta})f_2(\boldsymbol{\theta}) + c_2(\boldsymbol{\theta})g_2(\boldsymbol{\theta})$$
$$\boldsymbol{\theta} \in \mathbb{R}^2$$

where

$$h_1(\boldsymbol{\theta}) = \left|\frac{(-\theta_1 - 7)}{2} - \tanh(-\theta_2)\right|$$
$$h_2(\boldsymbol{\theta}) = \left|\frac{(-\theta_1 + 3)}{2} - \tanh(-\theta_2) + 2\right|$$
$$c_1(\theta) = \max(\tanh\left(\frac{\theta_2}{2}\right), 0)$$
$$c_2(\theta) = \max(\tanh\left(\frac{-\theta_2}{2}\right), 0)$$
$$f_1(\boldsymbol{\theta}) = \log\max\left(h_1(\boldsymbol{\theta}), 5\cdot 10^{-6}\right) + 6$$
$$f_2(\boldsymbol{\theta}) = \log\max\left(h_2(\boldsymbol{\theta}), 5\cdot 10^{-6}\right) + 6$$
$$g_1(\boldsymbol{\theta}) = \frac{(-\theta - 7)^2 + 0.1(-\theta_2 - 8)^2}{10} - 20$$
$$g_2(\boldsymbol{\theta}) = \frac{(-\theta + 7)^2 + 0.1(-\theta_2 - 8)^2}{10} - 20$$

We perform minimization starting from five initial points: $[-8.5, 7.5], [0.0, 0.0], [9.0, 9.0], [-7.5, -0.5], [9, -1.0]$. We use Adam [1] optimizer with learning rate 10^{-3} and optimize for 35k iterations. We demonstrate that our method is able to converge to the optimums with varying pre-defined task weights in Fig. 2. For this purpose we explore a number of task convex combinations, such that $\mathcal{L}_0 = \alpha\mathcal{L}_1 + (1-\alpha)\mathcal{L}_2$

D. Implementation details

CITYSCAPES three-task. Following MGDA-UB training setup [6], we train PSPNet [9] model for 100 epochs using Adam optimizer with learning rate 10^{-4}. Train batch size is set to 8. Images from training set are resized into 512×256 resolution. We augment training set using random rotation and horizontal flips. The performance is averaged across 3 random initializations.

CITYSCAPES two-task. We follow CAGrad [2] training setup and train MTAN [4] model. Semantic labels are groupped into 7 classes. Batch size is set to 8, learning rate of Adam optimizer is set to 10^{-4}. Models are trained for 200 epochs and learning rate is halved after 100 epochs. The performance is averaged over last 10 epochs and 3 random seeds.

Figure 2. Comparison of MTL optimization methods on synthetic two-task benchmark [2, 5]. We explore convergence of various methods with varying pre-defined task weights. Methods that guarantee only Pareto-front convergence (such as IMTL [3] and NashMTL [5]) fail to achieve global optimum (defined by ★) and converge to an arbitrary Pareto-front solution with unknown task balance. Unlike previous methods, our Aligned-MTL approach respects pre-defined task weights and converges to the global optimum for all task weights combinations and initialization points (●), except one extreme case. Moreover, our method provides stable and less noisy trajectories than other methods.

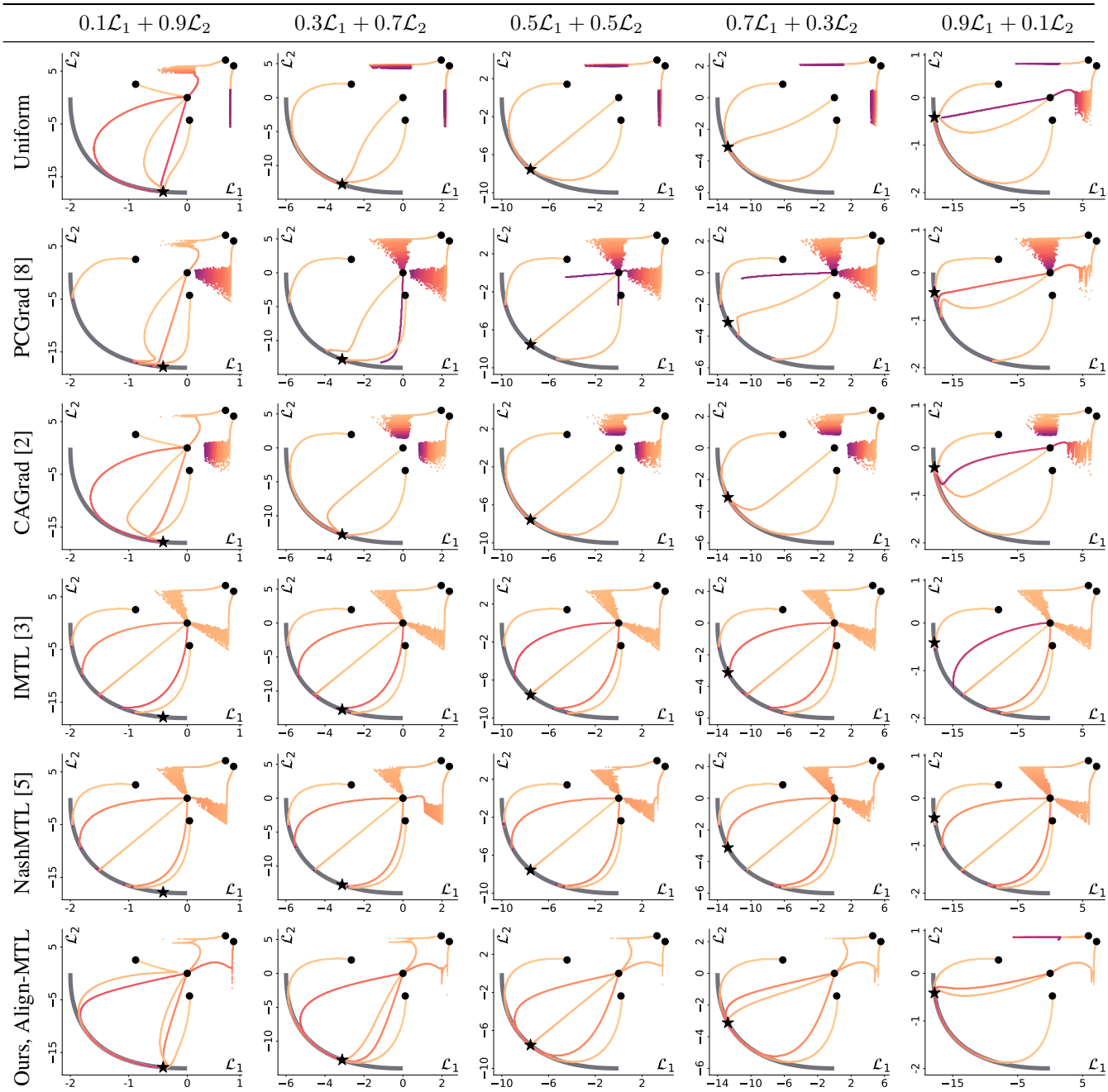
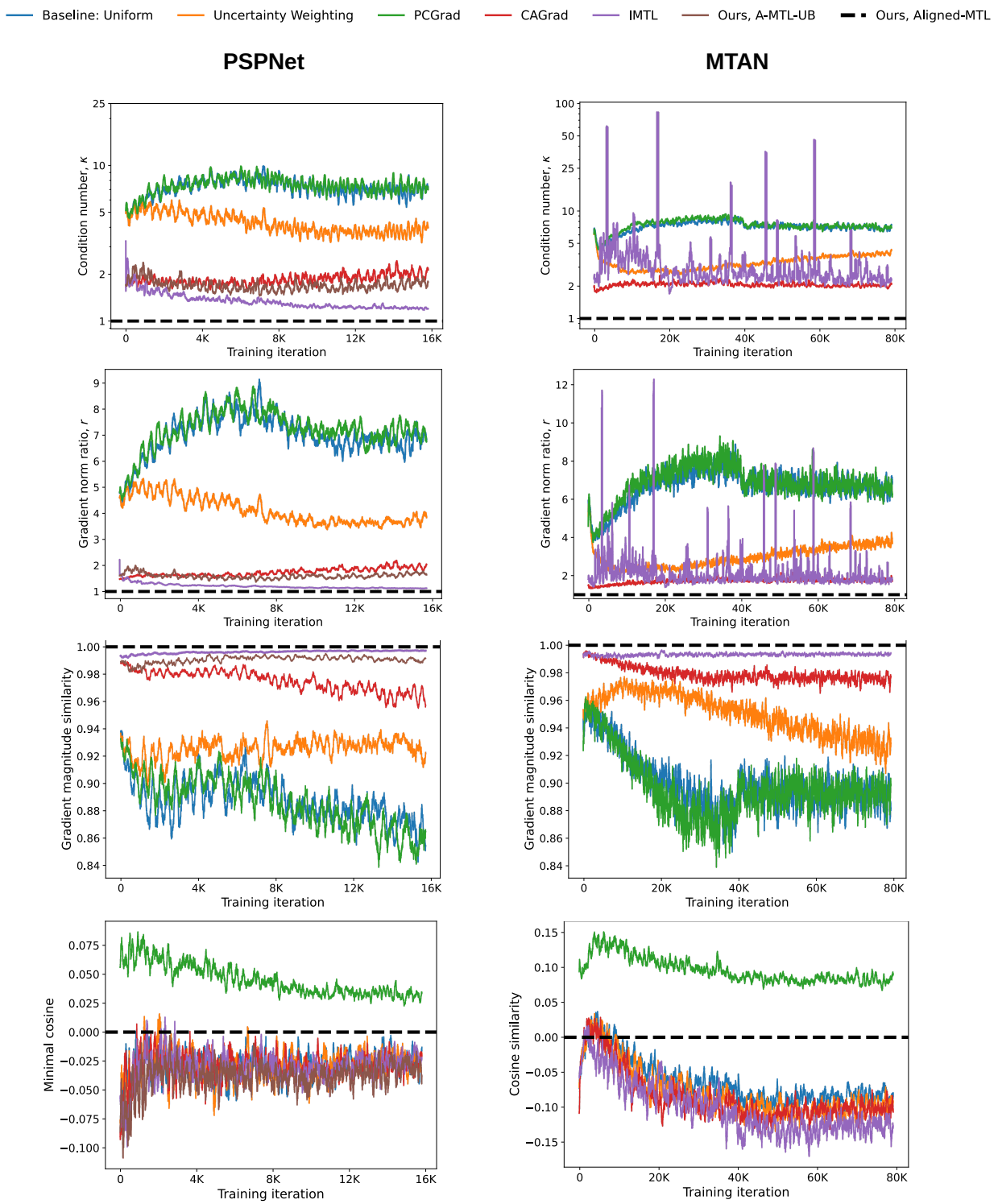


Figure 3. Empirical evaluation of a stability criterion. We plot a condition number, gradient magnitude similarity [8], minimal cosine between gradient pairs (conflicts) and maximum gradient norm ratio, *i.e.* $\max_{i \neq j}\{\|\boldsymbol{g}_i\|/\|\boldsymbol{g}_j\|\}$, during training of PSPNet [6, 9] and MTAN [4] on the NYUv2 benchmark. Unlike Cityscapes with three tasks (figure from the main paper), on NYUv2 gradients do not differ drastically in magnitudes but tend to have more conflicts (the cosine between gradients are negative, except for PCGrad). These figures indicate a high correlation between condition number, gradient norm ratios and gradient magnitude similarity. Our Aligned-MTL approach eliminates dominance ($\kappa = 1, r = 1, GMS = 1$) and conflicts ($\min_{i \neq j} cos(g_i, g_j) = 0$) by design.



NYUv2 three-task. [2, 4, 5] We train both PSPNet models [6, 9] and MTAN [4] models in our training setup with the same hyperparameters set. We use Adam [1] optimizer with learning rate 10^{-4}. Models are trained for 200 epochs and batch size 2. Images from training set are randomly scaled and cropped into 384×288 resolution. The performance is averaged across 3 random seeds.

Reinforcement learning. We follow CAGrad [2] and use the implementation originally proposed and developed by [7]. The execution config was adapded from CAGrad [2]. The global evaluation pipeline is similar to previous works [2, 5]. The performance is averaged over 10 random seeds.

References

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.

[2] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 18878–18890. Curran Associates, Inc., 2021.

[3] Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *ICLR*, 2021.

[4] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. In *CVPR*, pages 1871–1880, 2019.

[5] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 16428–16446. PMLR, 2022.

[6] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 527–538. Curran Associates, Inc., 2018.

[7] Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9767–9779. PMLR, 18–24 Jul 2021.

[8] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 5824–5836. Curran Associates, Inc., 2020.

[9] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 6230–6239, 2017.