Supplementary Materials

A Experiments

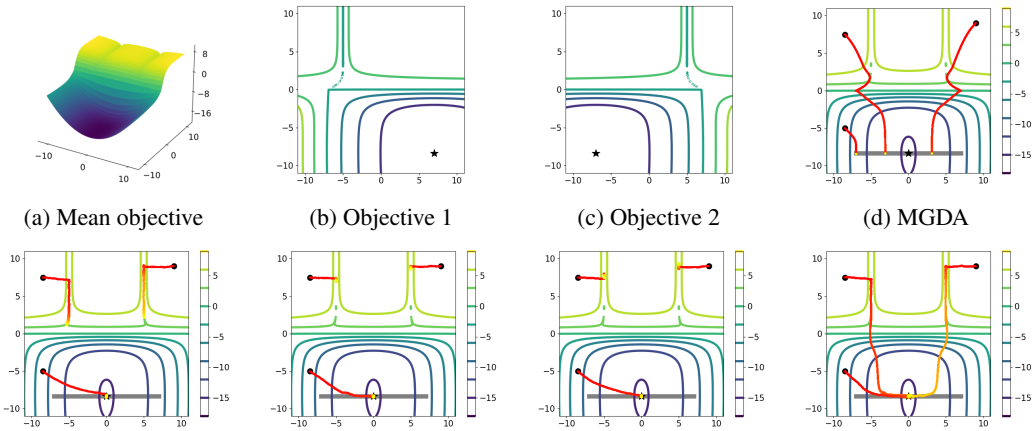
A.1 Toy Example

To demonstrate our proposed method can achieve better or comparable performance under stochastic settings, we provide an empirical study on the two-objective toy example used in CAGrad [9]. The two objectives $L_1(x)$ and $L_2(x)$ shown in Figure 1 are defined on $x = (x_1, x_2)^\top \in \mathbb{R}^2$,

$$\begin{aligned} L_1(x) &= f_1(x)g_1(x) + f_2(x)h_1(x) \\ L_2(x) &= f_1(x)g_2(x) + f_2(x)h_2(x), \end{aligned}$$

where the functions are given by

$$\begin{aligned} f_1(x) &= \max\big(\tanh(0.5x_2), 0\big) \\ f_2(x) &= \max\big(\tanh(-0.5x_2), 0\big) \\ g_1(x) &= \log\Big(\max\big(|0.5(-x_1 - 7) - \tanh(-x_2)|, 0.000005\big)\Big) + 6 \\ g_2(x) &= \log\Big(\max\big(|0.5(-x_1 + 3) - \tanh(-x_2) + 2|, 0.000005\big)\Big) + 6 \\ h_1(x) &= \big((-x_1 + 7)^2 + 0.1(-x_1 - 8)^2\big)/10 - 20 \\ h_2(x) &= \big((-x_1 - 7)^2 + 0.1(-x_1 - 8)^2\big)/10 - 20. \end{aligned}$$



(a) Mean objective (b) Objective 1 (c) Objective 2 (d) MGDA

(e) GD (f) PCGrad (g) CAGrad (h) SDMGrad(Ours)

Figure 1: A two-objective toy example.

We choose 3 initializations

$$x_0 \in \{(-8.5, 7.5), (-8.5, 5), (9, 9)\}$$

for different methods and visualize the optimization trajectories in Figure 1. The starting point of every trajectory in Figure 1d-Figure 1h is given by the $\bullet$ symbol, and the color of every trajectory changes gradually from red to yellow. The gray line illustrates the Pareto front, and the $\star$ symbol denotes the global optimum. To simulate the stochastic setting, we add zero-mean Gaussian noise to the gradient of each objective for all the methods except MGDA. We adopt Adam optimizer with learning rate of 0.002 and 70000 iterations for each run. As shown, GD can get stuck due to the dominant gradient of a specific objective, which stops progressing towards the Pareto front. PCGrad and CAGrad can also fail to converge to the Pareto front in certain circumstances.

A.2 Consistency Verification

We conduct the experiment on the multi-task classification dataset Multi-Fashion+MNIST [45]. Each image contained in this dataset is constructed by overlaying two images randomly sampled from MNIST [46] and FashionMNIST [47] respectively. We adopt shrinked Lenet [48] as the shared base-encoder and a task-specific linear classification head for each task. We report the training losses obtained from different methods over 3 independent runs in Figure 2. As illustrated, the performance of SDMGrad with large λ is similar to GD, and the performance when λ is small resembles MGDA. With properly tuned λ, lower average training loss can be obtained. Generally, the results confirm the consistency of our formulation with the direction-oriented principle.

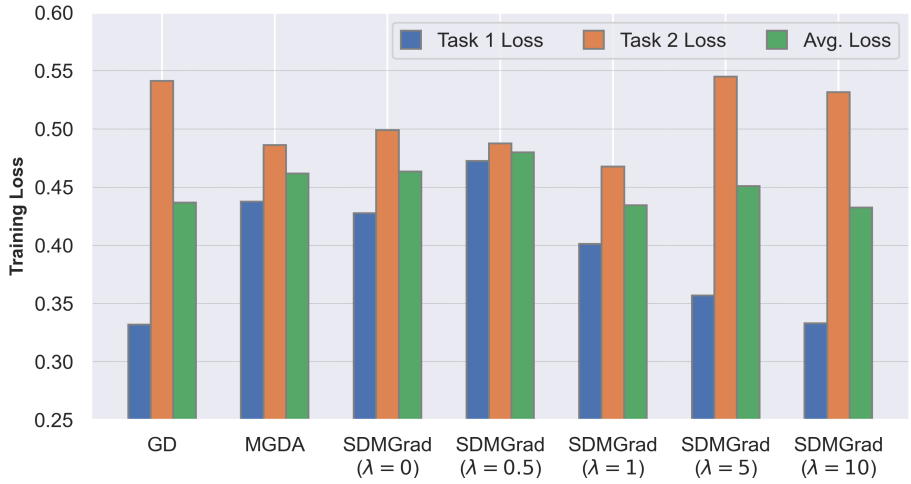


Figure 2: Consistency verification on Multi-Fashion+MNIST dataset.

The Multi-Fashion+MNIST[45] includes images constructed from FashionMNIST[47] and MNIST[46]. First, select one image from each dataset randomly, then transform the two images into a single image with one put in the top-left corner and the other in bottom-right corner. The dataset contains 120000 training images and 20000 test images. We use SGD optimizer with learning rate 0.001 and train for 100 epochs with batch size 256. We use multi-step scheduler with scale factor 0.1 to decay learning rate every 15 epochs. The projected gradient descent is performed with learning rate of 10 and momentum of 0.5 and 20 gradient descent steps are applied.

A.3 Supervised Learning

We implement the methods based on the library released by [10]. Following [9, 13, 10], we train our method for 200 epochs, using Adam optimizer with learning rate 0.0001 for the first 100 epochs and 0.00005 for the rest. The batch size for Cityscapes and NYU-v2 are 8 and 2 respectively. We compute the averaged test performance over the last 10 epochs as final performance measure. The inner projected gradient descent is performed with learning rate of 10 and momentum of 0.5 and 20 gradient descent steps are applied. The experiments on Cityscapes and NYU-v2 are run on RTX 3090 and Tesla V100 GPU, respectively. We also report additional experiment results over different λ and $S = 1$ in Table 5 and Table 6.

A.4 Reinforcement Learning

Following [9, 13, 10], we conduct the experiments based on MTRL codebase[49]. We train our method for 2 million steps with batch size of 1280. The inner projected gradient descent is performed with learning rate of 10 for MT10 benchmark and 20 gradient descent steps are applied. The method is evaluated once every 10000 steps and the best average test performance over 10 random seeds over the entire training process is reported. We search $\lambda \in \{0.1, 0.2, \cdots, 1.0\}$ for MT10 benchmark and the highest success rate is achieved when $\lambda = 0.6$. For our objective sampling strategy, the number of sampled objectives is a random variable obeying binomial distribution whose expectation is n.

Method	Segmentation		Depth		$\Delta m\% \downarrow$
	mIoU ↑	Pix Acc ↑	Abs Err ↓	Rel Err ↓	
STL	74.01	93.16	0.0125	27.77	
SDMGrad ($\lambda = 0.1$)	72.56	92.68	0.0156	40.89	18.65
SDMGrad ($\lambda = 0.2$)	74.79	93.30	0.0149	32.46	8.62
SDMGrad ($\lambda = 0.3$)	74.53	93.52	0.0137	34.01	**7.79**
SDMGrad ($\lambda = 0.4$)	75.10	93.48	0.0137	35.66	9.11
SDMGrad ($\lambda = 0.5$)	74.63	93.46	0.0131	38.99	11.09
SDMGrad ($\lambda = 0.6$)	74.42	93.22	0.0138	38.79	12.30
SDMGrad ($\lambda = 0.7$)	75.06	93.42	0.0158	39.98	17.24
SDMGrad ($\lambda = 0.8$)	74.99	93.40	0.0155	39.65	16.30
SDMGrad ($\lambda = 0.9$)	75.60	93.50	0.0134	43.52	15.39
SDMGrad ($\lambda = 1.0$)	74.50	93.47	0.0142	42.80	16.41
SDMGrad ($\lambda = 10$)	74.17	93.13	0.0154	41.77	18.36
SDMGrad ($\lambda = 0.3, S = 1$)	75.41	93.62	0.0139	38.83	12.22

Table 5: Addtional supervised learning experiments on Cityscapes dataset.

Method	Segmentation		Depth		Surface Normal					$\Delta m\% \downarrow$
	mIoU ↑	Pix Acc ↑	Abs Err ↓	Rel Err ↓	Angle Distance ↓		Within t° ↑			
					Mean	Median	11.25	22.5	30	
STL	38.30	63.76	0.6754	0.2780	25.01	19.21	30.14	57.20	69.15	
SDMGrad ($\lambda = 0.1$)	40.23	66.01	0.5360	0.2268	25.03	19.99	28.45	55.80	68.65	-3.86
SDMGrad ($\lambda = 0.2$)	39.23	65.67	0.5315	0.2189	25.13	20.02	28.12	55.71	68.46	-3.66
SDMGrad ($\lambda = 0.3$)	40.47	65.90	0.5225	0.2084	25.07	19.99	28.54	55.74	68.53	**-4.84**
SDMGrad ($\lambda = 0.4$)	40.68	66.53	0.5248	0.2199	25.21	20.01	27.69	55.72	68.58	-4.14
SDMGrad ($\lambda = 0.5$)	41.08	66.82	0.5184	0.2116	25.65	20.68	26.70	54.27	67.46	-3.33
SDMGrad ($\lambda = 0.6$)	41.20	66.86	0.5258	0.2175	25.85	21.03	26.47	53.51	66.82	-2.39
SDMGrad ($\lambda = 0.7$)	41.00	66.31	0.5224	0.2202	25.60	20.64	27.64	54.30	67.15	-3.16
SDMGrad ($\lambda = 0.8$)	39.88	66.13	0.5406	0.2266	26.20	21.57	25.67	52.33	65.65	-0.09
SDMGrad ($\lambda = 0.9$)	41.03	67.16	0.5314	0.2271	25.89	20.97	27.22	53.58	66.48	-2.17
SDMGrad ($\lambda = 1.0$)	39.94	66.27	0.5224	0.2155	26.51	21.95	25.15	51.54	64.94	-0.06
SDMGrad ($\lambda = 10$)	39.81	66.11	0.5352	0.2232	27.05	22.57	24.53	50.24	63.59	1.82
SDMGrad ($\lambda = 0.3, S = 1$)	39.63	65.43	0.5296	0.2140	25.66	20.83	27.18	53.93	67.05	-2.34

Table 6: Addtional supervised learning experiments on NYU-v2 dataset.

To compare with CAGrad-Fast[9], we choose $n = 4$ for MT10 benchmark. We cite the reported success rates of all baseline methods in Table 4, but independently run each experiment 5 times to calculate the average running time. All experiments on MT10 are run on RTX 2080Ti GPU. We also report addtional experiments results over $S = 1$ on MT10 in Table 7.

Method	Metaworld MT10	
	success (mean ± stderr)	time
SDMGrad	**0.84** ± 0.10	13.6
SDMGrad (S=1)	0.83 ± 0.05	11.2
SDMGrad-OS	0.82 ± 0.08	9.7
SDMGrad-OS (S=1)	0.80 ± 0.12	**6.8**

Table 7: Additonal reinforcement learning experiments on Metaworld MT10 benchmarks.

B Notations for Technical Proofs

In this part, we first summarize all the notations that we used in this paper in order to help readers understand. First, in multi-objective optimization, we have $K \geq 2$ different objectives and each of them has the loss function $L_i(\theta)$. Let g_i denote the gradient of objective i and g_0 denotes the target gradient. $w = (w_1, ..., w_K)^T \in \mathbb{R}^K$ *and* $\mathcal{W}$ denotes the probability simplex. Other useful notations are listed as below:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^m} \left\{ L_0(\theta) \triangleq \frac{1}{K} \sum_{i=1}^{K} L_i(\theta) \right\}, \quad g_0 = g_0(\theta) = G(\theta)\widetilde{w}$$

$$g_w = \sum_i w_i g_i \quad s.t. \quad \mathcal{W} = \{w : \sum_i w_i = 1 \ \ and \ \ w_i \geq 0\}$$

$$w_\lambda^* = \arg\min_{w\in\mathcal{W}} \frac{1}{2}\|g_w + \lambda g_0\|^2, \quad w^* = \arg\min_{w\in\mathcal{W}} \frac{1}{2}\|g_w\|^2, \quad w_t^* = \arg\min_{w\in\mathcal{W}} \frac{1}{2}\|G(\theta_t)w\|^2$$

$$w_{t,\lambda}^* = \arg\min_{w\in\mathcal{W}} F(w) = \arg\min_{w\in\mathcal{W}} \frac{1}{2}\|G(\theta_t)w + \lambda g_0(\theta_t)\|^2$$

$$\nabla_w F(w) = G(\theta_t)^T(G(\theta_t)w + \lambda g_0(\theta_t)), \nabla_w \widehat{F}(w) = G(\theta_t;\xi)^T(G(\theta_t,\xi')w + \lambda g_0(\theta_t,\xi')). \quad (11)$$

We use $\mathbb{E}[\cdot]_{A|B}$ to denote taking expectation over A conditioning on B and $\widetilde{\mathcal{O}}$ omits the order of $\log$.

C Detailed proofs for convergence analysis with nonconvex Objectives

We now provide some auxiliary lemmas for proving Proposition 1 and Theorem 1

Lemma 1. *Let d^* be the solution of*

$$\max_{d\in\mathbb{R}^m} \min_{i\in[K]} \langle g_i, d\rangle - \frac{1}{2}\|d\|^2 + \lambda\langle g_0, d\rangle,$$

then we have

$$d^* = g_{w_\lambda^*} + \lambda g_0.$$

In addition, w_λ^* is the solution of

$$\min_{w\in\mathcal{W}} \frac{1}{2}\|g_w + \lambda g_0\|^2.$$

Proof. First, it can be seen that

$$\begin{aligned}
&\max_{d\in\mathbb{R}^m} \min_{i\in[K]} \langle g_i, d\rangle - \frac{1}{2}\|d\|^2 + \lambda\langle g_0, d\rangle \\
&= \max_{d\in\mathbb{R}^m} \min_{w\in\mathcal{W}} \langle \sum_i w_i g_i, d\rangle - \frac{1}{2}\|d\|^2 + \lambda\langle g_0, d\rangle \\
&= \max_{d\in\mathbb{R}^m} \min_{w\in\mathcal{W}} {g_w}^T d - \frac{1}{2}\|d\|^2 + \lambda\langle g_0, d\rangle. \quad (12)
\end{aligned}$$

Noting that the problem is concave w.r.t. d and convex w.r.t w and using the Von Neumann-Fan minimax theorem [50], we can exchange the min and max problems without changing the solution. Then, we can solve the following equivalent problem.

$$\min_{w\in\mathcal{W}} \max_{d\in\mathbb{R}^m} {g_w}^T d - \frac{1}{2}\|d\|^2 + \lambda\langle g_0, d\rangle \quad (13)$$

Then by fixing w, we have $d^* = g_w + \lambda g_0$. Substituting this solution to the eq. (13) and rearranging the equation, we turn to solve the following problem.

$$\min_{w\in\mathcal{W}} \frac{1}{2}\|g_w + \lambda g_0\|^2.$$

Let w_λ^* be the solution of the above problem, and hence the final updating direction $d^* = g_{w_\lambda^*} + \lambda g_0$. Then, the proof is complete. □

Lemma 2. *Suppose Assumption 2-3 are satisfied. According to the definition of $g_0(\theta)$ in eq. (11), we have the following inequalities,*

$$\|g_0(\theta)\| \leq C_g, \quad \mathbb{E}[\|g_0(\theta;\xi) - g_0(\theta)\|^2] \leq K\sigma_0^2.$$

Proof. Based on the definitions, we have

$$\|g_0(\theta)\| = \|G(\theta)\widetilde{w}\| \leq C_g,$$

where the inequality follows from the fact that $\|\widetilde{w}\| \leq 1$ and Assumption 3. Then, we have

$$\mathbb{E}_\xi[\|g_0(\theta;\xi) - g_0(\theta)\|^2] \leq \mathbb{E}_\xi[\|G(\theta;\xi) - G(\theta)\|^2] \leq K\sigma_0^2$$

where $\sigma_0^2 = \max_i \sigma_i^2$ and the proof is complete. □

Lemma 3. *Suppose Assumptions 2, 3 are satisfied and recall that* $F(w) = \frac{1}{2}\|G(\theta_t)w + \lambda g_0(\theta_t)\|^2$ *is a convex function. Let* $w_\lambda^* = \arg\min_{w\in\mathcal{W}} \frac{1}{2}\|g_w + \lambda g_0\|^2$ *and set step size* $\beta_{t,s} = c/\sqrt{s}$ *where* $c > 0$ *is a constant. Then for any* $S > 1$*, it holds that,*

$$\mathbb{E}[\|\nabla_w \widehat{F}(w)\|] \leq C_1,$$
$$\mathbb{E}[\|G(\theta_t)w_S + \lambda g_0(\theta_t)\|^2 - \|G(\theta_t)w_\lambda^* + \lambda g_0(\theta_t)\|^2] \leq (\frac{2}{c} + 2cC_1)\frac{2 + log(S)}{\sqrt{S}}$$

where $C_1 = \sqrt{8(K\sigma_0^2 + C_g^2)^2 + 8\lambda^2(K\sigma_0^2 + C_g^2)^2} = \mathcal{O}(K + \lambda K)$, $\nabla_w \widehat{F}(w) = G(\theta_t;\xi)^T(G(\theta_t;\xi')w + \lambda g_0(\theta_t;\xi'))$.

Proof. This lemma mostly follows from Theorem 2 in [51]. However, we did not take that $\mathbb{E}[\|\nabla_w \widehat{F}(w)\|]$ is bounded by a constant as an assumption. Therefore, we first provide a bound for it in our method. Based on the definition in Equation (11), $\nabla_w \widehat{F}(w) = G(\theta_t;\xi)^T(G(\theta_t;\xi')w + \lambda g_0(\theta_t;\xi'))$. According to the fact that $\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}$, we have

$$\begin{aligned}\mathbb{E}[\|\nabla_w \widehat{F}(w)\|] &\leq \sqrt{\mathbb{E}[\|\nabla_w \widehat{F}(w)\|^2]} = \sqrt{\mathbb{E}[\|G(\theta_t;\xi)^T(G(\theta_t;\xi')w + \lambda g_0(\theta_t;\xi'))\|^2]} \\ &\overset{(i)}{\leq} \sqrt{2\mathbb{E}[\underbrace{\|G(\theta_t;\xi)^T G(\theta_t;\xi')w\|^2}_{A} + \lambda^2 \underbrace{\|G(\theta_t;\xi)^T g_0(\theta_t;\xi'))\|^2}_{B}]}, \end{aligned} \tag{14}$$

where (i) follows from the Young's inequality. Next, we provide bounds for $\mathbb{E}[A]$ and $\mathbb{E}[B]$, separately:

$$\begin{aligned}\mathbb{E}[A] \overset{(i)}{\leq}& \mathbb{E}[\|(G(\theta_t;\xi)^T - G(\theta_t)^T + G(\theta_t)^T)(G(\theta_t;\xi') - G(\theta_t) + G(\theta_t))\|^2] \\ =& \mathbb{E}[\|(G(\theta_t;\xi)^T - G(\theta_t)^T)(G(\theta_t;\xi') - G(\theta_t)) + (G(\theta_t;\xi)^T - G(\theta_t)^T)G(\theta_t) \\ &+ G(\theta_t)^T(G(\theta_t;\xi') - G(\theta_t)) + G(\theta_t)^T G(\theta_t)\|^2] \\ \overset{(ii)}{\leq}& 4\mathbb{E}[\|G(\theta_t;\xi)^T - G(\theta_t)^T\|^2\|G(\theta_t;\xi') - G(\theta_t)\|^2 + \|G(\theta_t;\xi)^T - G(\theta_t)^T\|^2\|G(\theta_t)\|^2 \\ &+ \|G(\theta_t)^T\|^2\|(G(\theta_t;\xi') - G(\theta_t)\|^2 + \|G(\theta_t)^T G(\theta_t)\|^2] \\ \overset{(iii)}{\leq}& 4K^2\sigma_0^4 + 8K\sigma_0^2 C_g^2 + 4C_g^4 = 4(K\sigma_0^2 + C_g^2)^2, \end{aligned} \tag{15}$$

where (i) follows from Cauchy–Schwarz inequality and $w \in \mathcal{W}$ where $\mathcal{W}$ is the simplex, (ii) follows from Young's inequality and (iii) follows from Assumption 2 and Assumption 3. Then for term B, we have,

$$\begin{aligned}\mathbb{E}[B] =& \mathbb{E}[\|(G(\theta_t;\xi)^T - G(\theta_t)^T + G(\theta_t)^T)(g_0(\theta_t;\xi') - g_0(\theta_t) + g_0(\theta_t))\|^2] \\ \overset{(i)}{\leq}& 4\mathbb{E}[\|(G(\theta_t;\xi)^T - G(\theta_t)^T)(g_0(\theta_t;\xi') - g_0(\theta_t))\|^2 + \|(G(\theta_t;\xi)^T - G(\theta_t)^T)g_0(\theta_t)\|^2 \\ &+ \|G(\theta_t)^T(g_0(\theta_t;\xi') - g_0(\theta_t))\|^2 + \|G(\theta_t^T)g_0(\theta_t)\|^2] \\ \overset{(ii)}{\leq}& 4K^2\sigma_0^4 + 8K\sigma_0^2 C_g^2 + 4C_g^4 = 4(K\sigma_0^2 + C_g^2)^2, \end{aligned} \tag{16}$$

where (i) follows from Young's inequality, (ii) follows from Assumption 3 and Lemma 2. Then substituting eq. (15) and eq. (16) into eq. (14), we can obtain,

$$\mathbb{E}[\|\nabla_w \widehat{F}(w)\|] \leq \sqrt{8(K\sigma_0^2 + C_g^2)^2 + 8\lambda^2(K\sigma_0^2 + C_g^2)^2} = C_1.$$

Meanwhile, since $\mathbb{E}[\|\nabla_w \widehat{F}(w)\|] \leq C_1$, $\sup_{w,w'} \|w - w'\| \leq 1$ and by choosing step size $\beta_s = c/\sqrt{s}$ where $c > 0$ is a constant, we can obtain the following inequality from Theorem 2 in [51]:

$$\mathbb{E}[F(w_S) - F(w_\lambda^*)] \leq (\frac{1}{c} + cC_1)\frac{2 + log(S)}{\sqrt{S}} \tag{17}$$

Then after multiplying by 2 on both sides, the proof is complete. □

C.1 Proof of Proposition 1

CA distance. Now we show the upper bound for the distance to CA direction. Recall that we define the CA distance as $\|\mathbb{E}_{\zeta,w_{t,S}|\theta_t}[G(\theta_t;\zeta)w_{t,S}+\lambda g_0(\theta_t;\zeta)]-G(\theta_t)w^*_{t,\lambda}-\lambda g_0(\theta_t)\|$.

Proof. Based on the Jensen's inequality, we have

$$
\begin{aligned}
&\|\mathbb{E}_{\zeta,w_{t,S}|\theta_t}[G(\theta_t;\zeta)w_{t,S}+\lambda g_0(\theta_t;\zeta)]-G(\theta_t)w^*_{t,\lambda}-\lambda g_0(\theta_t)\|^2\\
&\quad\le \mathbb{E}_{w_{t,S}|\theta_t}\left[\left\|\mathbb{E}_{\zeta}[G(\theta_t;\zeta)w_{t,S}+\lambda g_0(\theta_t;\zeta)]-G(\theta_t)w^*_{t,\lambda}-\lambda g_0(\theta_t)\right\|^2\right]\\
&\quad\overset{(i)}{=}\mathbb{E}[\|G(\theta_t)w_{t,S}-G(\theta_t)w^*_{t,\lambda}\|^2]\\
&\quad=\mathbb{E}[\|G(\theta_t)w_{t,S}+\lambda g_0(\theta_t)-G(\theta_t)w^*_{t,\lambda}-\lambda g_0(\theta_t)\|^2]\\
&\quad=\mathbb{E}[\|G(\theta_t)w_{t,S}+\lambda g_0(\theta_t)\|^2+\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2\\
&\qquad-2\mathbb{E}\langle G(\theta_t)w_{t,S}+\lambda g_0(\theta_t),G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\rangle]\\
&\quad=\mathbb{E}[\|G(\theta_t)w_{t,S}+\lambda g_0(\theta_t)\|^2+\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2]\\
&\qquad-2\mathbb{E}[\langle G(\theta)w_{t,S},G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\rangle]-2\mathbb{E}[\langle\lambda g_0(\theta_t),G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\rangle]\\
&\quad\overset{(ii)}{\le}\mathbb{E}[\|G(\theta_t)w_{t,S}+\lambda g_0(\theta_t)\|^2+\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2]\\
&\qquad-2\mathbb{E}[\langle G(\theta)w^*_{t,\lambda},G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\rangle]-2\mathbb{E}[\langle\lambda g_0(\theta_t),G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\rangle]\\
&\quad=\mathbb{E}[\|G(\theta_t)w_{t,S}+\lambda g_0(\theta_t)\|^2+\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2]\\
&\qquad-2\mathbb{E}[\langle G(\theta)w^*_{t,\lambda}+\lambda g_0(\theta_t),G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\rangle]\\
&\quad=\mathbb{E}[\|G(\theta_t)w_{t,S}+\lambda g_0(\theta_t)\|^2-\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2]\\
&\quad\overset{(iii)}{\le}(\frac{2}{c}+2cC_1)\frac{2+log(S)}{\sqrt{S}}
\end{aligned}
\tag{18}
$$

where (i) omits the subscript of taking expectation over $w_{t,S}$ conditioning on θ_t, (ii) follows from optimality condition that

$$
\langle w,G(\theta_t)^T(G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t))\rangle\ge\langle w^*_{t,\lambda},G(\theta_t)^T(G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t))\rangle. \tag{19}
$$

(iii) follows from Lemma 3 whenwe choose $\beta_{t,s}=c/\sqrt{s}$ where c is a constant. Then take the square root on both sides, the proof is complete. □

C.2 Proof of Theorem 1

Theorem 5 (Restatement of Theorem 1)**.** *Suppose Assumptions 1-3 are satisfied. Set* $\alpha_t=\alpha=\Theta((1+\lambda)^{-1}K^{-\frac{1}{2}}T^{-\frac{1}{2}})$, $\beta_{t,s}=c/\sqrt{s}$ *where* c *is a constant, and* $S=\Theta((1+\lambda)^{-2}T^2)$. *The outputs of the proposed SDMGrad algorithm satisfy*

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2]=\widetilde{\mathcal{O}}((1+\lambda^2)K^{\frac{1}{2}}T^{-\frac{1}{2}}).
$$

Proof. Recall that $d=G(\theta_t;\zeta)w_{t,S}+\lambda g_0(\theta_t;\zeta)$. According to Assumption 1, we have for any i,

$$
L_i(\theta_{t+1})+\lambda L_0(\theta_{t+1})\le L_i(\theta_t)+\lambda L_0(\theta_t)+\alpha_t\langle g_i(\theta_t)+\lambda g_0(\theta_t),-d\rangle+\frac{l'_{i,1}\alpha_t^2}{2}\|d\|^2. \tag{20}
$$

where $l'_{i,1}=l_{i,1}+\lambda\max_i l_{i,1}=\Theta(1+\lambda)$. Then we bound the second and third terms separately on the right-hand side (RHS). First, for the second term, conditioning on θ_t and taking expectation, we have

$$
\begin{aligned}
&\mathbb{E}[\langle g_i(\theta_t)+\lambda g_0(\theta_t),-G(\theta_t;\zeta)w_{t,S}-\lambda g_0(\theta_t;\zeta)\rangle|\theta_t]\\
&\quad=\mathbb{E}[\langle g_i(\theta_t)+\lambda g_0(\theta_t),-G(\theta_t)w_{t,S}-\lambda g_0(\theta_t)\rangle|\theta_t]\\
&\quad=\mathbb{E}[\langle g_i(\theta_t)+\lambda g_0(\theta_t),G(\theta_t)w^*_{t,\lambda}-G(\theta_t)w_{t,S}\rangle-\langle g_i(\theta_t)+\lambda g_0(\theta_t),G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\rangle|\theta_t]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(i)}{\leq}\mathbb{E}[(l_i+\lambda C_g)\|G(\theta_t)w^*_{t,\lambda}-G(\theta_t)w_{t,S}\||\theta_t]-\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2|\theta_t]\\
&\overset{(ii)}{\leq}(l_i+\lambda C_g)\sqrt{\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda}-G(\theta_t)w_{t,S}\|^2|\theta_t]}-\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2|\theta_t]\\
&\overset{(iii)}{\leq}(l_i+\lambda C_g)\sqrt{(\frac{2}{c}+2cC_1)\frac{2+log(S)}{\sqrt{S}}}-\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2|\theta_t] \qquad (21)
\end{aligned}
$$

where (i) follows from Cauchy-Schwarz inequality and optimality condition in eq. (19), (ii) follows from the fact that $\mathbb{E}[X]\leq\sqrt{\mathbb{E}[X^2]}$ and (iii) follows from eq. (18).

Then for the third term,

$$
\begin{aligned}
\mathbb{E}[\|d\|^2]=&\mathbb{E}[\|G(\theta_t;\zeta)w_{t,S}+\lambda g_0(\theta_t;\zeta)\|^2]\\
=&\mathbb{E}[\|G(\theta_t;\zeta)w_{t,S}-G(\theta_t)w_{t,S}+G(\theta_t)w_{t,S}+\lambda g_0(\theta_t;\zeta)-\lambda g_0(\theta_t)+\lambda g_0(\theta_t)\|^2]\\
\overset{(i)}{\leq}&4\mathbb{E}[\|G(\theta_t;\zeta)-G(\theta_t)\|^2]+4\mathbb{E}[\|G(\theta_t)\|^2]+4\lambda^2\mathbb{E}[\|g_0(\theta_t;\zeta)-g_0(\theta_t)\|^2]\\
&+4\lambda^2\mathbb{E}[\|g_0(\theta_t)\|^2]\\
\overset{(ii)}{\leq}&\underbrace{4K\sigma_0^2+4C_g^2+4\lambda^2K\sigma_0^2+4\lambda^2C_g^2}_{C_2} \qquad (22)
\end{aligned}
$$

where (i) follows from Young's inequality, and (ii) follows from Assumption 3 and Lemma 2. Note that $C_2=\mathcal{O}(K+K\lambda^2)$. Then taking expectation on eq. (20), substituting eq. (21) and eq. (22) into it, and unconditioning on θ_t, we have

$$
\begin{aligned}
&\mathbb{E}[L_i(\theta_{t+1})+\lambda L_0(\theta_{t+1})]\\
&\quad\leq\mathbb{E}[L_i(\theta_t)+\lambda L_0(\theta_t)]+\alpha_t\mathbb{E}[\langle g_i(\theta_t)+\lambda g_0(\theta_t),-d\rangle]+\frac{l'_{i,1}\alpha_t^2}{2}\mathbb{E}[\|d\|^2]\\
&\quad\leq\mathbb{E}[L_i(\theta_t)+\lambda L_0(\theta_t)]+\alpha_t(l_i+\lambda C_g)\sqrt{(\frac{2}{c}+2cC_1)\frac{2+log(S)}{\sqrt{S}}}\\
&\qquad-\alpha_t\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2]+\frac{l'_{i,1}\alpha_t^2}{2}C_2 \qquad (23)
\end{aligned}
$$

Then, choosing $\alpha_t=\alpha$, and rearranging the above inequality, we have

$$
\begin{aligned}
\alpha\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2]\leq&\mathbb{E}[L_i(\theta_t)+\lambda L_0(\theta_t)-L_i(\theta_{t+1})-\lambda L_0(\theta_{t+1})]\\
&+\alpha(l_i+\lambda C_g)\sqrt{(\frac{2}{c}+2cC_1)\frac{2+log(S)}{\sqrt{S}}}+\frac{l'_{i,1}\alpha^2}{2}C_2.
\end{aligned}
$$

Telescoping over $t\in[T]$ in the above inequality yields

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2]\\
&\quad\leq\frac{1}{\alpha T}\mathbb{E}[L_i(\theta_0)-\inf L_i(\theta)+\lambda(L_0(\theta_0)-\inf L_0(\theta))]+\frac{l'_{i,1}\alpha}{2}C_2\\
&\qquad+(l_i+\lambda C_g)\sqrt{(\frac{2}{c}+2cC_1)\frac{2+log(S)}{\sqrt{S}}},
\end{aligned}
$$

If we choose $\alpha=\Theta((1+\lambda)^{-1}K^{-\frac{1}{2}}T^{-\frac{1}{2}})$ and $S=\Theta((1+\lambda)^{-2}T^2)$, we have

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda}+\lambda g_0(\theta_t)\|^2]=\widetilde{\mathcal{O}}((1+\lambda^2)K^{\frac{1}{2}}T^{-\frac{1}{2}}),
$$

where $\widetilde{\mathcal{O}}$ means the order of $logT$ is omitted. The proof is complete. □

C.3 Proof of Corollary 1

Proof. Since $\lambda > 0$ and $g_0(\theta_t) = G(\theta_t)\widetilde{w}$, we have

$$\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda} + \lambda g_0(\theta_t)\|^2] = (1+\lambda)^2\mathbb{E}[\|G(\theta_t)w'\|^2] \geq (1+\lambda)^2\mathbb{E}[\|G(\theta_t)w^*_t\|^2]$$

where $w' = \frac{1}{1+\lambda}(w^*_{1,t,\lambda} + \lambda\widetilde{w}_1, w^*_{2,t,\lambda} + \lambda\widetilde{w}_2, ..., w^*_{K,t,\lambda} + \lambda\widetilde{w}_K)^T$ such that $w' \in \mathcal{W}$. According to parameter selection in Theorem 1 and by choosing a constant λ, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w^*_t\|^2] = \widetilde{\mathcal{O}}(K^{\frac{1}{2}}T^{-\frac{1}{2}}). \tag{24}$$

To achieve an ϵ-accurate Pareto stationary point, it requires $T = \widetilde{\mathcal{O}}(K\epsilon^{-2})$ and each objective requires $\widetilde{\mathcal{O}}(K^3\epsilon^{-6})$ samples in ξ (ξ') and $\widetilde{\mathcal{O}}(K\epsilon^{-2})$ samples in ζ, respectively. Meanwhile, according to the choice of S and T, we have the following result for CA distance,

$$\|\mathbb{E}_{\zeta,w_{t,S}|\theta_t}[G(\theta_t,\zeta)w_{t,S} + \lambda g_0(\theta_t;\zeta)] - G(\theta_t)w^*_{t,\lambda} - \lambda g_0(\theta_t)\| = \widetilde{\mathcal{O}}(\sqrt{\frac{K}{T}}) = \widetilde{\mathcal{O}}(\epsilon) \tag{25}$$

Remark. Our algorithm with a constant λ helps mitigate gradient conflict and it guarantees an $\epsilon-$accurate Pareto stationary point and the CA distance takes the order of $\widetilde{O}(\epsilon)$ simultaneously. □

C.4 Proof of Corollary 2

Proof. According to the inequality $\|a+b-b\|^2 \leq 2\|a+b\|^2 + 2\|b\|^2$, we have

$$\begin{aligned}\lambda^2\|g_0(\theta_t)\|^2 \leq& 2\|G(\theta_t)w^*_{t,\lambda} + \lambda g_0(\theta_t)\|^2 + 2\|G(\theta_t)w^*_{t,\lambda}\|^2\\ \leq& 2\|G(\theta_t)w^*_{t,\lambda} + \lambda g_0(\theta_t)\|^2 + 2C_g^2\end{aligned}$$

where the last inequality follows from Assumption 3. Then we take the expectation on the above inequality and sum up it over $t \in [T]$ such that

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|g_0(\theta_t)\|^2] \leq& \frac{2}{\lambda^2 T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w^*_{t,\lambda} + \lambda g_0(\theta_t)\|^2] + \frac{C_g^2}{\lambda^2}\\ =& \widetilde{\mathcal{O}}((\lambda^{-2}+1)K^{\frac{1}{2}}T^{-\frac{1}{2}} + \lambda^{-2}),\end{aligned}$$

where the last inequality follows from Theorem 1. If we choose $\lambda = \Theta(T^{\frac{1}{2}})$, then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|g_0(\theta_t)\|^2] = \widetilde{\mathcal{O}}(K^{\frac{1}{2}}T^{-\frac{1}{2}}).$$

To achieve an ϵ-accurate stationary point, it requires $T = \widetilde{\mathcal{O}}(K\epsilon^{-2})$ and each objective requires $\widetilde{\mathcal{O}}(K^2\epsilon^{-4})$ samples in ξ (ξ') and $\widetilde{\mathcal{O}}(K\epsilon^{-2})$ samples in ζ, respectively. Meanwhile, according to the choice of λ, S and T, we have the following result for CA distance,

$$\|\mathbb{E}_{\zeta,w_{t,S}|\theta_t}[G(\theta_t,\zeta)w_{t,S} + \lambda g_0(\theta_t;\zeta)] - G(\theta_t)w^*_{t,\lambda} - \lambda g_0(\theta_t)\| = \widetilde{\mathcal{O}}(\sqrt{\frac{K(1+\lambda)^2}{T}}) = \widetilde{\mathcal{O}}(\sqrt{K})$$

Remark. With an increasing λ, our algorithm approaches GD and it has a faster convergence rate to the stationary point. However, the CA distance takes the order of $\widetilde{O}(\sqrt{K})$. □

C.5 Proof of Theorem 2

Now we provide the convergence analysis with nonconvex objectives with objective sampling.

Theorem 6 (Restatement of Theorem 2)**.** *Suppose Assumptions 1-3 are satisfied. Set* $\gamma = \frac{K}{n}$, $\alpha_t = \alpha = \Theta((1+\lambda^2)^{-\frac{1}{2}}\gamma^{-\frac{1}{2}}K^{-\frac{1}{2}}T^{-\frac{1}{2}})$, $\beta_{t,s} = c/\sqrt{s}$ *and* $S = \Theta((1+\lambda)^{-2}\gamma^{-2}T^2)$. *Then by choosing a constant* λ*, the iterates of the proposed SDMGrad-OS algorithm satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w^*_t\|^2] = \widetilde{\mathcal{O}}(K^{\frac{1}{2}}\gamma^{\frac{1}{2}}T^{-\frac{1}{2}}).$$

Proof. Recall that updating direction for θ_t is $d' = \frac{K}{n} H(\theta_t; \zeta, \widetilde{S}) w_{t,S} + \frac{K}{n} \lambda h_0(\theta_t; \zeta, \widetilde{S})$. Similarly, we have

$$L_i(\theta_{t+1}) + \lambda L_0(\theta_{t+1}) \leq L_i(\theta_t) + \lambda L_0(\theta_t) + \alpha_t \langle g_i(\theta_t) + \lambda g_0(\theta_t), -d' \rangle + \frac{l'_{i,1} \alpha_t^2}{2} \|d'\|^2. \quad (26)$$

Then for the inner product term on the RHS of eq. (26), conditioning on θ_t and taking expectation, we have

$$\begin{aligned}\mathbb{E}[\langle g_i(\theta_t) + \lambda g_0(\theta_t), -d' \rangle | \theta_t] &= \mathbb{E}[\langle g_i(\theta_t) + \lambda g_0(\theta_t), -G(\theta_t) w_{t,S} - \lambda g_0(\theta_t) \rangle | \theta_t] \\ &\leq (l_i + \lambda C_g)\Big(\sqrt{(\frac{2}{c} + 2cC_1)\frac{2 + log(S)}{\sqrt{S}}}\Big) - \mathbb{E}[\|G(\theta_t) w^*_{t,\lambda} + \lambda g_0(\theta_t)\|^2 | \theta_t], \quad (27)\end{aligned}$$

where the last inequality follows from eq. (21). Then following the same step as in eq. (22), we can bound the last term on the RHS of eq. (26) as

$$\mathbb{E}[\|d'\|^2] \leq \underbrace{4\gamma^2(1 + \lambda^2)(n\sigma_0^2 + C_g^2)}_{C'_2}. \quad (28)$$

Then taking expectation on eq. (26), substituting eq. (27) and eq. (28) into it and unconditioning on θ_t, we have

$$\begin{aligned}\mathbb{E}[L_i(\theta_{t+1}) + \lambda L_0(\theta_{t+1})] \leq &\mathbb{E}[L_i(\theta_t) + \lambda L_0(\theta_t)] + \alpha_t (l_i + \lambda C_g)\Big(\sqrt{(\frac{2}{c} + 2cC_1)\frac{2 + log(S)}{\sqrt{S}}}\Big) \\ &- \alpha_t \mathbb{E}\|G(\theta_t) w^*_{t,\lambda} + \lambda g_0(\theta_t)\|^2 + \frac{l'_{i,1} \alpha_t^2}{2} C'_2\end{aligned}$$

Then choosing $\alpha_t = \alpha$, telescoping the above inequality over $t \in [T]$, and rearranging the terms, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|G(\theta_t) w^*_{t,\lambda} + \lambda g_0(\theta_t)\|^2] \leq &\frac{1}{\alpha T} \mathbb{E}[L_i(\theta_0) - \inf L_i(\theta) + \lambda(L_0(\theta_0) - \inf L_0(\theta))] + \frac{l'_{i,1}\alpha}{2} C'_2 \\ &+ (l_i + \lambda C_g)\Big(\sqrt{(\frac{2}{c} + 2cC_1)\frac{2 + log(S)}{\sqrt{S}}}\Big)\end{aligned}$$

If we choose $\alpha = \Theta((1 + \lambda^2)^{-\frac{1}{2}} \gamma^{-\frac{1}{2}} K^{-\frac{1}{2}} T^{-\frac{1}{2}})$, and $S = \Theta((1 + \lambda)^{-2} \gamma^{-2} T^2)$, we can get $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|G(\theta_t) w^*_{t,\lambda} + \lambda g_0(\theta_t)\|^2] = \widetilde{\mathcal{O}}((1 + \lambda^2) K^{\frac{1}{2}} \gamma^{\frac{1}{2}} T^{-\frac{1}{2}})$. Furthermore, by choosing λ as constant and following the same step as in Appendix C.3, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|G(\theta_t) w^*_t\|^2] = \widetilde{\mathcal{O}}(K^{\frac{1}{2}} \gamma^{\frac{1}{2}} T^{-\frac{1}{2}}).$$

To achieve an ϵ-accurate Pareto stationary point, it requires $T = \widetilde{\mathcal{O}}(K\gamma\epsilon^{-2})$. In this case, each objective requires a similar number of samples $\widetilde{\mathcal{O}}(K^3\gamma\epsilon^{-6})$ in ξ (ξ') and $\widetilde{\mathcal{O}}(K\gamma\epsilon^{-2})$ samples in ζ, respectively. As far as we know, this is the first provable objective sampling strategy for stochastic multi-objective optimization. □

D Lower sample complexity but higher CA distance

When we do not have requirements on CA distance, we can have a much lower sample complexity. In Algorithm 1, the update process for w is to reduce the CA distance, which increases the sample complexity. Thus, we will set $S = 1$ to make Algorithm 1 more sample-efficient. In addition, we will use $w_{t+1} = w_{t,1}$ and β_t instead of $\beta_{t,s}$ in Algorithm 1 for simplicity. The following proof is mostly motivated by Theorem 3 in [14].

D.1 Proof of Theorem 3

Theorem 7 (Restatement of Theorem 3)**.** *Suppose Assumptions 1-3 are satisfied and $S = 1$. Set $\alpha_t = \alpha = \Theta(K^{-\frac{1}{2}}T^{-\frac{1}{2}})$, $\beta_t = \beta = \Theta(K^{-1}T^{-\frac{1}{2}})$ and λ as constant. The iterates of the proposed SDMGrad algorithm satisfy,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w_t^*\|^2] = \mathcal{O}(KT^{-\frac{1}{2}}).$$

Proof. Now we define a new function, with a fixed weight $w \in \mathcal{W}$,

$$l'(\theta_t) = L(\theta_t)w + \lambda L_0(\theta_t). \tag{29}$$

For this new function, we have

$$\begin{aligned}
l'(\theta_{t+1}) \leq & l'(\theta_t) + \alpha_t\langle G(\theta_t)w + \lambda g_0(\theta_t), -d\rangle + \frac{l_1'\alpha_t^2}{2}\|d\|^2 \\
= & l'(\theta_t) + \alpha_t\langle G(\theta_t)w + \lambda g_0(\theta_t), -G(\theta_t;\zeta)w_{t+1} - \lambda g_0(\theta_t;\zeta)\rangle + \frac{l_1'\alpha_t^2}{2}\|d\|^2
\end{aligned}$$

where $l_1' = \max_i l_{i,1} + \lambda l_{i,1}$. Then taking expectations over ζ on both sides and rearranging the inequality, we have

$$\begin{aligned}
\mathbb{E}[l'(\theta_{t+1})] - \mathbb{E}[l'(\theta_t)] \leq & \alpha_t\mathbb{E}[\langle G(\theta_t)w + \lambda g_0(\theta_t), -G(\theta_t)w_{t+1} - \lambda g_0(\theta_t)\rangle] + \frac{l_1'\alpha_t^2}{2}\mathbb{E}[\|d\|^2] \\
= & -\alpha_t\mathbb{E}[\langle G(\theta_t)w + \lambda g_0(\theta_t), G(\theta_t)w_{t+1} - G(\theta_t)w_t\rangle] \\
& - \alpha_t\mathbb{E}[\langle G(\theta_t)w + \lambda g_0(\theta_t), G(\theta_t)w_t + \lambda g_0(\theta_t)\rangle] + \frac{l_1'\alpha_t^2}{2}\mathbb{E}[\|d\|^2] \\
= & -\alpha_t\mathbb{E}[\langle G(\theta_t)w + \lambda g_0(\theta_t), G(\theta_t)w_{t+1} - G(\theta_t)w_t\rangle] \\
& - \alpha_t\mathbb{E}[\langle G(\theta_t)w - G(\theta_t)w_t, G(\theta_t)w_t + \lambda g_0(\theta_t)\rangle] \\
& - \alpha_t\mathbb{E}[\|G(\theta_t)w_t + \lambda g_0(\theta_t)\|^2] + \frac{l_1'\alpha_t^2}{2}\mathbb{E}[\|d\|^2] \\
\overset{(i)}{\leq} & \alpha_t \underbrace{\mathbb{E}[\|(G(\theta_t)w + \lambda g_0(\theta_t))^T G(\theta_t)\|\|w_t - w_{t+1}\|]}_{C} \\
& + \alpha_t \underbrace{\mathbb{E}[\langle G(\theta_t)w_t - G(\theta_t)w, G(\theta_t)w_t + \lambda g_0(\theta_t)\rangle]}_{D} \\
& - \alpha_t\mathbb{E}[\|G(\theta_t)w_t + \lambda g_0(\theta_t)\|^2] + \frac{l_1'\alpha_t^2}{2}\mathbb{E}[\|d\|^2],
\end{aligned} \tag{30}$$

where (i) follows from Cauchy-Schwarz inequlaity. Then we provide bound for term C and term D, respectively. For term C,

$$\begin{aligned}
& \mathbb{E}[\|(G(\theta_t)w + \lambda g_0(\theta_t))^T G(\theta_t)\|\|w_t - w_{t+1}\|] \\
= & \beta_t\mathbb{E}[\|(G(\theta_t)w + \lambda g_0(\theta_t))^T G(\theta_t)\|\|G(\theta_t;\xi)^T(G(\theta_t;\xi')w_t + \lambda g_0(\theta;\xi'))\|] \\
\leq & \beta_t\mathbb{E}[\|(G(\theta_t)w + \lambda g_0(\theta_t))^T G(\theta_t)\|(\|G(\theta_t;\xi)^T(G(\theta_t;\xi')w_t\| + \lambda\|G(\theta_t;\xi)g_0(\theta;\xi')\|)] \\
\leq & \beta_t(1+\lambda)^2C_g^2(K\sigma_0 + C_g)^2 = \beta_t C_3,
\end{aligned} \tag{31}$$

where $C_3 = \mathcal{O}((1+\lambda)^2K^2)$. Then for term D, we first follow the non-expansive property of projection onto the convex set,

$$\begin{aligned}
\|w_{t+1} - w\|^2 \leq & \|w_t - \beta_t G(\theta_t;\xi)^T(G(\theta_t;\xi')w_t + \lambda g_0(\theta_t;\xi')) - w\|^2 \\
= & \|w_t - w\|^2 - 2\beta_t\langle w_t - w, G(\theta_t;\xi)^T(G(\theta_t;\xi')w_t + \lambda g_0(\theta_t;\xi'))\rangle \\
& + \beta_t^2\|G(\theta_t;\xi)^T(G(\theta_t;\xi')w_t + \lambda g_0(\theta_t;\xi'))\|^2
\end{aligned}$$

Then taking expectation on the above inequality, we can obtain,

$$\mathbb{E}[\|w_{t+1} - w\|^2] \leq \mathbb{E}[\|w_t - w\|^2] - 2\beta_t\mathbb{E}[\langle w_t - w, G(\theta_t;\xi)^T(G(\theta_t;\xi')w_t + \lambda g_0(\theta_t;\xi'))\rangle]$$

$$
\begin{aligned}
&+\beta_t^2\mathbb{E}[\|G(\theta_t;\xi)^T(G(\theta_t;\xi')w_t+\lambda g_0(\theta_t;\xi'))\|^2]\\
\leq&\mathbb{E}[\|w_t-w\|^2]-2\beta_t\mathbb{E}[\langle w_t-w,G(\theta_t)^T(G(\theta_t)w_t+\lambda g_0(\theta_t))\rangle]+\beta_t^2C_1^2,
\end{aligned}
$$

where the last inequality follows from Lemma 3. Then by rearranging the above inequality, we can obtain,

$$
\mathbb{E}[\langle w_t-w,G(\theta_t)^T(G(\theta_t)w_t+\lambda g_0(\theta_t))\rangle]\leq\frac{1}{2\beta_t}\mathbb{E}[\|w_t-w\|^2-\|w_{t+1}-w\|^2]+\frac{\beta_t}{2}C_1^2 \tag{32}
$$

Then substituting eq. (31) and eq. (32) into eq. (30), we can obtain,

$$
\begin{aligned}
\mathbb{E}[l'(\theta_{t+1})-l'(\theta_t)]\leq&\alpha_t\beta_tC_3+\frac{\alpha_t}{2\beta_t}\mathbb{E}[\|w_t-w\|^2-\|w_{t+1}-w\|^2]+\frac{\alpha_t\beta_t}{2}C_1^2\\
&-\alpha_t\mathbb{E}[\|G(\theta_t)w_t+\lambda g_0(\theta_t)\|^2]+\frac{l_1'\alpha_t^2}{2}\mathbb{E}[\|d\|^2]\\
\overset{(i)}{\leq}&\alpha_t\beta_tC_3+\frac{\alpha_t}{2\beta_t}\mathbb{E}[\|w_t-w\|^2-\|w_{t+1}-w\|^2]+\frac{\alpha_t\beta_t}{2}C_1^2\\
&-\alpha_t\mathbb{E}[\|G(\theta_t)w_t+\lambda g_0(\theta_t)\|^2]+\frac{l_1'\alpha_t^2}{2}C_2
\end{aligned} \tag{33}
$$

Then we take $\alpha_t=\alpha$ and $\beta_t=\beta$ as constants, telescope and rearrange the above inequality,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w_t+\lambda g_0(\theta_t)\|^2]\leq&\frac{1}{\alpha T}\mathbb{E}[l'(\theta_0)-l'(\theta_T)]+\frac{1}{2\beta T}\mathbb{E}[\|w_0-w\|^2-\|w_T-w\|^2]\\
&+\beta(C_3+\frac{C_1^2}{2})+\frac{l_1'\alpha}{2}C_2\\
\overset{(i)}{\leq}&\mathcal{O}(\frac{1}{\alpha T}+\alpha K+\frac{1}{\beta T}+\beta K^2),
\end{aligned} \tag{34}
$$

where (i) follows from that we choose λ as a constant. If we choose $\alpha=\Theta(K^{-\frac{1}{2}}T^{-\frac{1}{2}})$ and $\beta=\Theta(K^{-1}T^{-\frac{1}{2}})$, we can get $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w_t+\lambda g_0(\theta_t)\|^2]=\mathcal{O}(KT^{-\frac{1}{2}})$. Furthermore, following the same steps as in Appendix C.3, we have

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w_t^*\|^2]=\mathcal{O}(KT^{-\frac{1}{2}}).
$$

To achieve an ϵ-accurate Pareto stationary point, it requires $T=\mathcal{O}(K^2\epsilon^{-2})$. In this case, each objective requires a similar number of samples $\mathcal{O}(K^2\epsilon^{-2})$ in $\xi(\xi')$ and ζ, respectively. □

Convergence under objective sampling. We next analyze the convergence of SDMGrad-OS.

Theorem 8 (Restatement of Theorem 4)**.** *Suppose Assumptions 1-3 are satisfied and $S=1$. Set $\gamma=\frac{K}{n}$, $\alpha_t=\alpha=\Theta(K^{-\frac{1}{2}}\gamma^{-\frac{1}{2}}T^{-\frac{1}{2}})$, $\beta_t=\beta=\Theta(K^{-1}\gamma^{-1}T^{-\frac{1}{2}})$ and λ as a constant. The iterates of the proposed SDMGrad-OS algorithm satisfy,*

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w_t^*\|^2]=\mathcal{O}(K\gamma T^{-\frac{1}{2}}).
$$

Proof. In SDMGrad-OS, the vector for updating θ_t is $d'=\frac{K}{n}H(\theta_t;\zeta,\widetilde{S})w_{t+1}+\frac{\lambda K}{n}h_0(\theta_t;\zeta,\widetilde{S})$. Using the same function defined in eq. (29), we have

$$
l'(\theta_{t+1})\leq l'(\theta_t)+\alpha_t\langle G(\theta_t)w+\lambda g_0(\theta_t),-d'\rangle+\frac{l_1'\alpha_t^2}{2}\|d'\|^2.
$$

Then by taking expectation over ζ and $\widetilde{S}$, we have

$$
\mathbb{E}[l'(\theta_{t+1})-l'(\theta_t)]\leq\alpha_t\mathbb{E}[\langle G(\theta_t)w+\lambda g_0(\theta_t),-G(\theta_t)w_{t+1}+\lambda g_0(\theta_t)\rangle]+\frac{l'\alpha_t^2}{2}\mathbb{E}[\|d'\|^2]
$$

$$
\begin{aligned}
=&\alpha_t E[\langle G(\theta_t)w + \lambda g_0(\theta_t), G(\theta_t)(w_t - w_{t+1})\rangle] \\
&+ \alpha_t \mathbb{E}[\langle G(\theta_t)w_t - G(\theta_t)w, G(\theta_t)w_t + \lambda g_0(\theta_t)\rangle] \\
&- \mathbb{E}[\|G(\theta_t)w_t + \lambda g_0(\theta_t)\|^2] + \frac{l_1'\alpha_t^2}{2}\mathbb{E}[\|d'\|^2] \\
\leq&\alpha_t \mathbb{E}[\|(G(\theta_t)w + \lambda g_0(\theta_t))^T G(\theta_t)\|\|w_t - w_{t+1}\|] \\
&+ \alpha_t \mathbb{E}[\langle G(\theta_t)w_t - G(\theta_t)w, G(\theta_t)w_t + \lambda g_0(\theta_t)\rangle] \\
&- \mathbb{E}[\|G(\theta_t)w_t + \lambda g_0(\theta_t)\|^2] + \frac{l_1'\alpha_t^2}{2}\mathbb{E}[\|d'\|^2] \qquad (35)
\end{aligned}
$$

Then following the same steps in eq. (31) and eq. (32), we can obtain,

$$
\begin{aligned}
\mathbb{E}[l'(\theta_{t+1}) - l'(\theta_t)] \leq&\alpha_t\beta_t C_3' + \frac{\alpha_t}{2\beta_t}\mathbb{E}[\|w_t - w\|^2 - \|w_{t+1} - w\|^2] + \frac{\alpha_t\beta_t}{2}C_1'^2 \\
&-\alpha_t \mathbb{E}[\|G(\theta_t)w_t + \lambda g_0(\theta_t)\|^2] + \frac{l_1'\alpha_t^2}{2}C_2', \qquad (36)
\end{aligned}
$$

where $C_1'^2 = 4\gamma^4(1+\lambda^2)(n\sigma_0^2+C_g)^2$, $C_2' = 4\gamma^2(1+\lambda^2)(n\sigma_0^2+C_g^2)$, and $C_3' = \gamma^2(1+\lambda)^2C_g^2(n\sigma_0^2+C_g)^2$. Then we take $\alpha_t = \alpha$ and $\beta_t = \beta$ as constants and telescope the above inequality,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w_t + \lambda g_0(\theta_t)\|^2] \leq&\frac{1}{\alpha T}\mathbb{E}[l'(\theta_0) - l'(\theta_T)] + \frac{1}{2\beta T}\mathbb{E}[\|w_0 - w\|^2 - \|w_T - w\|^2] \\
&+ \beta(C_3' + \frac{C_1'^2}{2}) + \frac{l_1'\alpha}{2}C_2' \\
\overset{(i)}{\leq}&\mathcal{O}(\frac{1}{\alpha T} + \alpha\gamma K + \frac{1}{\beta T} + \beta\gamma^2K^2), \qquad (37)
\end{aligned}
$$

where (i) follows from that we choose λ as constant. Similarly, if we choose $\alpha = \Theta(K^{-\frac{1}{2}}\gamma^{-\frac{1}{2}}T^{-\frac{1}{2}})$ and $\beta = \Theta(K^{-1}\gamma^{-1}T^{-\frac{1}{2}})$, we can get $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w_t + \lambda g_0(\theta_t)\|^2] = \mathcal{O}(K\gamma T^{-\frac{1}{2}})$. Furthermore, following the same step as in Appendix C.3, we have

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|G(\theta_t)w_t^*\|^2] = \mathcal{O}(K\gamma T^{-\frac{1}{2}}).
$$

To achieve an ϵ-accurate Pareto stationary point, it requires $T = \mathcal{O}(\gamma^2K^2\epsilon^{-2})$. In this case, each objective requires a similar number of samples $\mathcal{O}(\gamma^2K^2\epsilon^{-2})$ in $\xi(\xi')$ and ζ, respectively. □