

## 482 A Problem formulations

### 483 A.1 Convex problems

484 The following formulations share:  $y \in \mathbb{R}^n$ ,  $Q \in \mathbb{S}_{++}^n$ ,  $p \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n_{\text{eq}} \times n}$ ,  $x \in \mathbb{R}^{n_{\text{eq}}}$ ,  $L, U \in \mathbb{R}^n$ .

$$\begin{aligned} \text{QP: } \min_{L \leq y \leq U} \quad & \frac{1}{2} y^\top Q y + p^\top y \\ \text{s.t. } & Ay = x, \quad Gy \leq h, \end{aligned}$$

485 where  $G \in \mathbb{R}^{n_{\text{ineq}} \times n}$ ,  $h \in \mathbb{R}^{n_{\text{ineq}}}$ .

$$\begin{aligned} \text{QCQP: } \min_{L \leq y \leq U} \quad & \frac{1}{2} y^\top Q y + p^\top y \\ \text{s.t. } & Ay = x, \quad y^\top H_i y + g_i^\top y \leq h_i, \end{aligned}$$

486 where  $H_i \in \mathbb{S}_{++}^n$ ,  $g_i \in \mathbb{R}^n$ ,  $h_i \in \mathbb{R}$  for  $i = 1, \dots, n_{\text{ineq}}$ .

$$\begin{aligned} \text{SOCP: } \min_{L \leq y \leq U} \quad & \frac{1}{2} y^\top Q y + p^\top y \\ \text{s.t. } & Ay = x, \quad \|G_i y + h_i\|_2 \leq c_i^\top y + d_i, \end{aligned}$$

487 where  $G_i \in \mathbb{R}^{m \times n}$ ,  $h_i \in \mathbb{R}^m$ ,  $c_i \in \mathbb{R}^n$ ,  $d_i \in \mathbb{R}$  for  $i = 1, \dots, n_{\text{ineq}}$ .

### 488 A.2 Nonconvex problems

489 We modify the convex problems by adding element-wise sine and cosine functions.

$$\begin{aligned} \text{Nonconvex QP: } \min_{L \leq y \leq U} \quad & \frac{1}{2} y^\top Q y + p^\top \sin(y) \\ \text{s.t. } & Ay = x, \quad G \sin(y) \leq h \cos(x), \end{aligned}$$

490

$$\begin{aligned} \text{Nonconvex QCQP: } \min_{L \leq y \leq U} \quad & \frac{1}{2} y^\top Q y + p^\top \sin(y) \\ \text{s.t. } & Ay = x, \quad y^\top H_i y + g_i^\top \cos(y) \leq h_i, \end{aligned}$$

491

$$\begin{aligned} \text{Nonconvex SOCP: } \min_{L \leq y \leq U} \quad & \frac{1}{2} y^\top Q y + p^\top \sin(y) \\ \text{s.t. } & Ay = x, \quad \|G_i \cos(y) + h_i\|_2 \leq c_i^\top y + d_i. \end{aligned}$$

### 492 A.3 Nonsmooth nonconvex problems

493 Compared to the (smooth) convex problems, we further add an  $\ell_2$ -norm regularization term to the  
494 objective function.

$$\begin{aligned} \text{Nonsmooth nonconvex QP: } \min_{L \leq y \leq U} \quad & \frac{1}{2} y^\top Q y + p^\top \sin(y) + \lambda \|y\|_2 \\ \text{s.t. } & Ay = x, \quad G \sin(y) \leq h \cos(x), \end{aligned}$$

495

$$\begin{aligned} \text{Nonsmooth nonconvex QCQP: } \min_{L \leq y \leq U} \quad & \frac{1}{2} y^\top Q y + p^\top \sin(y) + \lambda \|y\|_2 \\ \text{s.t. } & Ay = x, \quad y^\top H_i y + g_i^\top \cos(y) \leq h_i, \end{aligned}$$

496

$$\begin{aligned} \text{Nonsmooth nonconvex SOCP: } \min_{L \leq y \leq U} \quad & \frac{1}{2} y^\top Q y + p^\top \sin(y) + \lambda \|y\|_2 \\ \text{s.t. } & Ay = x, \quad \|G_i \cos(y) + h_i\|_2 \leq c_i^\top y + d_i. \end{aligned}$$

## B More experiment results

### B.1 Smooth convex problems

Table B.1 presents detailed results for smooth convex problems. Figure B.1 plots FSNet’s runtime on 2000 test instances as a function of batch size. Thanks to GPU parallelism, the runtime decreases monotonically as batch size increases, so it is recommended to use the largest batch size possible to maximize speedup. Figure B.2 reports CPU utilization for the parallel solver runs. Across all problems, CPU usage remains near 100% throughout execution, showing that we fully exploit available CPU resources. This indicates that our runtime comparisons between NN-based methods and the solver are fair.

Method	Equality Vio.	Inequality Vio.	Optimality Gap (%)		Runtime (s)
	Mean (Max)	Mean (Max)	Mean	Min (Max)	Batch (Sequential)
Convex QP: $n = 100, n_{eq} = 100, n_{ineq} = 100$					
Solver	1.3e-13±1.7e-16 (1.9e-13±5.5e-15)	2.4e-14±2.44e-16 (5.4e-14±4.19e-16)	—	—	3.802±0.045 (140.328±4.585)
Reduced Solver	1.2e-6±7.4e-8 (1.2e-6±7.4e-8)	2.1e-4±1.1e-5 (7.1e-5±1.0e-5)	2.5e-4±1.3e-5	-1.7e-7±2.3e-7 (5.3e-3±3.9e-4)	3.777±0.103 (127.105±5.372)
Projection (qpth)	9.7e-14±3.5e-16 (1.4e-13±3.4e-15)	1.2e-14±4.6e-16 (4.1e-13±2.2e-15)	3.7e-3±1.7e-3	2.2e-4±1e-4 (0.02±5.3e-3)	13.71±1.07 (225.063±43.585)
FSNet	1.3e-4±1.1e-5 (7.7e-4±1.9e-4)	1.6e-6±4.4e-7 (7.2e-5±3.2e-5)	0.015±0.002	5.7e-3±9.7e-4 (0.038±7.7e-4)	0.121±0.038 (180.389±22.505)
Penalty	0.219±5.8e-3 (5.4e-5±2.1e-5)	0.023±3.3e-3 (0.227±0.039)	-0.029±0.013	-0.074±0.013 (0.028±9.8e-3)	0.043±2.9e-3 (1.594±0.051)
Adaptive Penalty	0.221±7.6e-3 (0.371±2.9e-3)	0.022±1.6e-3 (0.211±0.019)	0.974±0.059	0.782±0.064 (1.213±0.074)	0.042±1.7e-3 (1.549±0.011)
DC3	4.9e-13±7.8e-16 (5.8e-13±6.8e-15)	0.133±0.011 (0.259±0.021)	-0.035±0.016	-0.087±0.019 (0.029±0.021)	0.058±3.5e-3 (53.151±0.558)
Convex QCQP: $n = 100, n_{eq} = 50, n_{ineq} = 50$					
Solver	8.4e-7±1.9e-8 (2.4e-5±5.8e-6)	1.5e-3±2.0e-5 (7.7e-3±0.00)	—	—	46.619±0.577 (1118.149±3.535)
FSNet	1.2e-4±1.2e-5 (1.5e-3±5.7e-4)	1.0e-6±2.2e-7 (6.4e-5±2.4e-5)	0.035±3.2e-3	7.8e-3±1.1e-3 (0.359±0.069)	0.448±0.016 (182.156±12.827)
Penalty	0.191±5.9e-3 (0.355±0.014)	0.031±0.01 (0.521±0.019)	0.038±0.062	-0.237±0.022 (0.397±0.2)	0.048±3.9e-3 (1.707±6.8e-3)
Adaptive Penalty	0.179±7.6±3 (0.333±7.6e-3)	0.028±2.1e-3 (0.497±0.075)	0.274±0.042	-0.119±0.067 (0.797±0.079)	0.047±2.8e-3 (1.703±0.02)
DC3	3.0e-13±4.9e-16 (3.8e-13±6.4e-15)	0.120±0.032 (0.602±0.072)	-0.018±0.037	-0.198±0.041 (0.259±0.039)	0.163±2.3e-3 (79.234±6.847)
Convex SOCP: $n = 100, n_{eq} = 50, n_{ineq} = 50$					
Solver	3.1e-14±4.2e-17 (2.2e-13±0.00)	1.3e-11 ±1.2e-11 (1.0e-8±8.8e-9)	—	—	332.109±2.186 (3702.308±9.56)
Reduced Solver	5.9e-7±1.3e-8 (4.8e-6±3.3e-7)	1.8e-4 ±4.1e-6 (2.0e-3±8.7e-5)	4.8e-5±2.1e-6	-1.2e-3±4.5e-5 (1.4e-3±1.1e-4)	330.143±1.687 (1895.989±4.592)
FSNet	1.3e-4±1.2e-5 (5.5e-4±9.9e-5)	1.7e-8±1.3e-8 (4.9e-6±1.5e-6)	0.159±5.0e-3	0.046±3.6e-3 (1.392±0.376)	0.295±8.0e-3 (208.399±9.935)
Penalty	0.098±5.2e-3 (0.139±3.6e-3)	0.015±3.2e-3 (0.309±0.053)	0.024±0.018	-0.235±0.107 (0.338±0.108)	0.049±3.4e-3 (1.919±0.236)
Adaptive Penalty	0.064±2.5e-3 (0.136±7.2e-3)	0.011±1.3e-3 (0.305±0.147)	0.256±0.126	-0.171±0.146 (0.808±0.359)	0.047±1.4e-3 (1.891±2.239)
DC3	7.8e-14±6.9e-17 (1.1e-13±9.8e-16)	0.029±6.0e-3 (0.325±0.049)	0.053±0.011	-0.103±9.5e-3 (0.304±0.064)	0.114±5.4e-3 (78.83±5.254)

Table B.1: Test results of on 2000 instances of smooth convex problems across 3 random seeds.

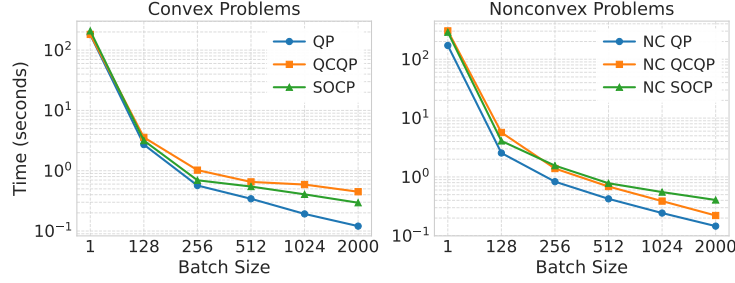


Figure B.1: Computational times of FSNet on 2000 instances of smooth convex and nonconvex (NC) problems with varying batch sizes.

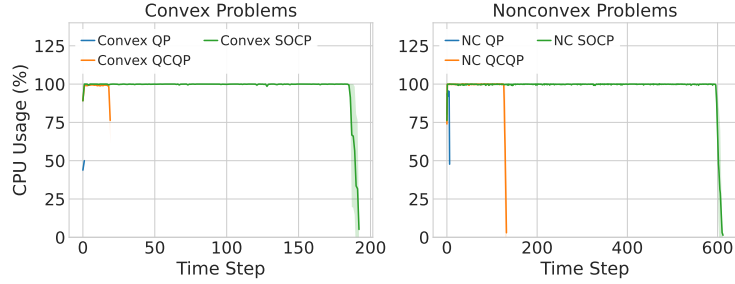


Figure B.2: CPU utilization over time for parallel solver runs on smooth convex and nonconvex problems.

## 506 B.2 Smooth nonconvex problems

507 Figure B.3 and Table B.2 present the detailed results for nonconvex problems. Figure B.1 and B.2 also  
 508 show the computational time of FSNet with different batch sizes and CPU utilization on nonconvex  
 509 cases, respectively. The patterns are the same as those observed in the smooth convex problems.

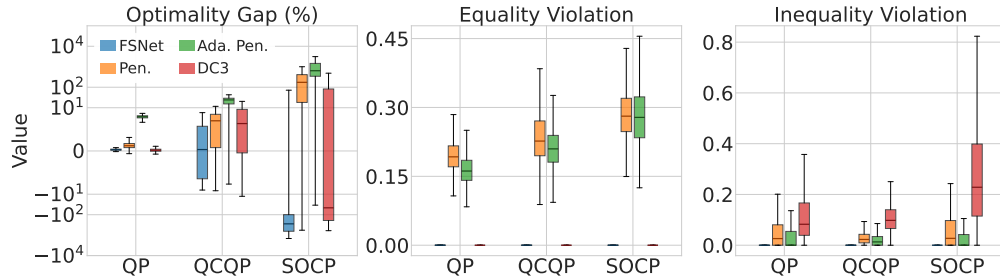


Figure B.3: Test results on 2000 instances of smooth nonconvex problems. FSNet consistently obtains near-zero constraint violations and small optimality gaps across the problem classes, whereas the baseline methods incur significant constraint violations.

Method	Equality Vio.	Inequality Vio.	Optimality Gap (%)		Runtime (s)
	Mean (Max)	Mean (Max)	Mean	Min (Max)	Batch (sequential)
<b>Nonconvex QP: <math>n = 100, n_{eq} = 100, n_{ineq} = 100</math></b>					
Solver	$5.5\text{e-}14 \pm 1.1\text{e-}16$ ( $8.3\text{e-}14 \pm 5.1\text{e-}15$ )	$1.8\text{e-}9 \pm 7.7\text{e-}11$ ( $1.6\text{e-}8 \pm 8.6\text{e-}10$ )	–	–	$12.699 \pm 0.692$ ( $470.846 \pm 0.991$ )
FSNet	$9.3\text{e-}5 \pm 1.9\text{e-}6$ ( $1.4\text{e-}3 \pm 2\text{e-}4$ )	$7.2\text{e-}7 \pm 8.7\text{e-}8$ ( $4.4\text{e-}5 \pm 1.1\text{e-}5$ )	$0.146 \pm 0.012$	$-0.341 \pm 0.24$ ( $8.885 \pm 0.592$ )	$0.146 \pm 0.038$ ( $171.226 \pm 10.671$ )
<b>Nonconvex QCQP: <math>n = 100, n_{eq} = 50, n_{ineq} = 50</math></b>					
Solver	$6.8\text{e-}14 \pm 1.6\text{e-}16$ ( $9.9\text{e-}14 \pm 3.5\text{e-}15$ )	$8.9\text{e-}9 \pm 2.1\text{e-}10$ ( $5.4\text{e-}8 \pm 1.6\text{e-}9$ )	–	–	$216.772 \pm 2.650$ ( $11169.226 \pm 102.22$ )
FSNet	$9.7\text{e-}5 \pm 3.9\text{e-}6$ ( $6.5\text{e-}6 \pm 2.9\text{e-}6$ )	$2.9\text{e-}6 \pm 2.4\text{e-}7$ ( $1.1\text{e-}4 \pm 1.4\text{e-}5$ )	$-0.668 \pm 0.650$	$-19.436 \pm 1.767$ ( $8.809 \pm 0.942$ )	$0.221 \pm 0.028$ ( $303.139 \pm 25.742$ )
<b>Nonconvex SOCP: <math>n = 100, n_{eq} = 50, n_{ineq} = 50</math></b>					
Solver	$8.3\text{e-}14 \pm 1.3\text{e-}16$ ( $1.3\text{e-}13 \pm 3.3\text{e-}15$ )	$9.9\text{e-}9 \pm 3.0\text{e-}10$ ( $7.6\text{e-}8 \pm 7.6\text{e-}10$ )	–	–	$1216.105 \pm 7.923$ ( $33765.581 \pm 170.102$ )
FSNet	$5.8\text{e-}5 \pm 3.5\text{e-}5$ ( $3.4\text{e-}3 \pm 1.4\text{e-}4$ )	$5.2\text{e-}7 \pm 1.5\text{e-}7$ ( $2.5\text{e-}4 \pm 1.6\text{e-}3$ )	$-2.3\text{e}3 \pm 0.7\text{e}3$	$-1.3\text{e}6 \pm 9.0\text{e}5$ ( $67.138 \pm 3.265$ )	$0.406 \pm 0.051$ ( $288.161 \pm 9.598$ )

Table B.2: Test results of FSNet on 2000 instances of smooth nonconvex problems.

### 510 B.3 Nonsmooth nonconvex problems

511 Table B.3 shows the detailed results of nonsmooth nonconvex problems (see Appendix A for formu-  
512 lations). As shown in this table, FSNet achieves near-zero constraint violations across nonsmooth  
513 nonconvex problems and achieves  $152\text{--}3316\times$  speedups in the batch setting and  $7\text{--}80\times$  speedups in  
514 the sequential setting compared to the traditional solver, which is also visualized in Figure B.5.

515 Notably, FSNet can attain lower-cost local solutions than the solver, as evidenced by the negative  
516 optimality gaps in Table B.3 and by the distributions of objective values in Figure B.4.

Method	Equality Vio.	Inequality Vio.	Optimality Gap (%)		Runtime (s)
	Mean (Max)	Mean (Max)	Mean	Min (Max)	Batch (sequential)
<b>Nonsmooth Nonconvex QP: <math>n = 100, n_{eq} = 100, n_{ineq} = 100</math></b>					
Solver	$5.4\text{e-}14 \pm 8.2\text{e-}17$ ( $7.9\text{e-}14 \pm 6.9\text{e-}16$ )	$5.3\text{e-}10 \pm 5.6\text{e-}12$ ( $1.2\text{e-}8 \pm 2.4\text{e-}10$ )	–	–	$19.161 \pm 0.566$ ( $1120.035 \pm 8.607$ )
FSNet	$9.0\text{e-}5 \pm 3.8\text{e-}6$ ( $1.5\text{e-}3 \pm 1.7\text{e-}4$ )	$6.3\text{e-}7 \pm 3.2\text{e-}8$ ( $4.2\text{e-}5 \pm 6.7\text{e-}6$ )	$0.109 \pm 8.5\text{e-}3$	$-1.081 \pm 0.275$ ( $8.128 \pm 0.593$ )	$0.126 \pm 0.024$ ( $156.553 \pm 4.761$ )
<b>Nonconvex QCQP: <math>n = 100, n_{eq} = 50, n_{ineq} = 50</math></b>					
Solver	$6.8\text{e-}14 \pm 2.4\text{e-}16$ ( $1.0\text{e-}13 \pm 1.4\text{e-}15$ )	$4.9\text{e-}9 \pm 3.7\text{e-}11$ ( $4.5\text{e-}8 \pm 3.5\text{e-}9$ )	–	–	$288.768 \pm 0.329$ ( $15092.158 \pm 6116.515$ )
FSNet	$9.1\text{e-}5 \pm 4.8\text{e-}6$ ( $6.5\text{e-}6 \pm 2.9\text{e-}6$ )	$2.9\text{e-}6 \pm 1.69\text{e-}7$ ( $1.1\text{e-}4 \pm 1.4\text{e-}5$ )	$-3.437 \pm 0.661$	$-27.299 \pm 2.573$ ( $9.466 \pm 0.919$ )	$0.539 \pm 0.018$ ( $247.819 \pm 37.821$ )
<b>Nonsmooth Nonconvex SOCP: <math>n = 100, n_{eq} = 50, n_{ineq} = 50</math></b>					
Solver	$7.8\text{e-}14 \pm 1.3\text{e-}16$ ( $1.2\text{e-}13 \pm 2.5\text{e-}15$ )	$3.4\text{e-}9 \pm 1.5\text{e-}10$ ( $5.9\text{e-}8 \pm 2.7\text{e-}9$ )	–	–	$1363.465 \pm 5.109$ ( $21642.764 \pm 81.569$ )
FSNet	$6.2\text{e-}5 \pm 9.4\text{e-}6$ ( $2.5\text{e-}3 \pm 5.2\text{e-}4$ )	$4.2\text{e-}7 \pm 1.2\text{e-}7$ ( $6.1\text{e-}5 \pm 6.0\text{e-}6$ )	$-1.2\text{e}3 \pm 9.43\text{e}2$	$-9.6\text{e}5 \pm 7.6\text{e}5$ ( $658.5 \pm 831.8$ )	$0.411 \pm 0.019$ ( $267.389 \pm 6.987$ )

Table B.3: Test results of FSNet on 2000 instances of nonsmooth nonconvex problems.

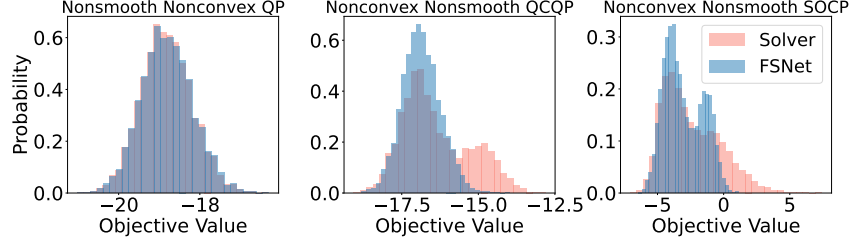


Figure B.4: Distribution of the objective values in nonsmooth nonconvex problems.

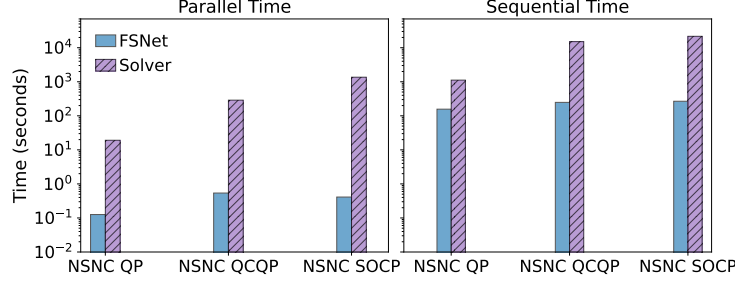


Figure B.5: Computational time on 2000 instances of nonsmooth nonconvex (NSNC) problems.

#### 517 B.4 FSNet with truncated unrolled differentiation

518 This section evaluates the impact of the truncation depth on the performance of FSNet (see Section  
 519 5). We set the maximum number of feasibility-seeking iterations to 50, and vary the truncation depth from 0 to 50. The results on the smooth convex problems are shown in Table B.4 and Figure B.6.

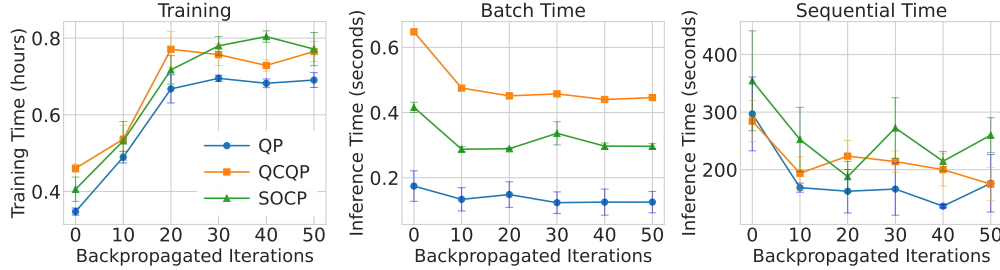


Figure B.6: Training time and batch inference time of FSNet with different truncation depths.

520

521 When  $K' = 0$ , no iterations of the feasibility-seeking procedure are included in the computational  
 522 graph, so it acts as a non-differentiable pass-through layer. This yields the smallest graph and therefore  
 523 the shortest training time, but incurs a large optimality gap, indicating the network cannot learn the  
 524 true solutions of Problem (1). As  $K'$  increases, the training is successful with small optimality gaps  
 525 (and near-zero constraint violations), yet training time grows due to the larger graph and additional  
 526 backpropagation cost (Figure B.6). Importantly, for  $K'$  between 10 and 50, both total constraint  
 527 violation and optimality gap remain essentially unchanged. This means that a modest value of  $K'$   
 528 is sufficient to render the bias caused by the gradient truncation negligible, which agrees with our  
 529 analysis in Appendix D.2. This suggests a practical tip that, rather than fully gradient-tracking all  
 530 iterations of the feasibility-seeking procedure, we only need to track a few of them while maintaining  
 531 high performance for FSNet and reducing the training time.

532 Since truncation affects only the backward pass during training, inference times are virtually identical  
 533 for all choices of  $K' \neq 0$ .

Tracked Iter. $K'$	Total vio. Mean (Max)	Optimality Gap (%) Mean (Max)	Runtime (s) Batch (Sequential)	Training time (h)
<b>Convex QP: <math>n = 100, n_{eq} = 100, n_{ineq} = 100</math></b>				
0	4.4e-5±1e-6 (4.2e-4±4.2e-5)	14.143±0.024 (16.314±0.148)	0.174±0.046 (296.889±64.220)	0.348±0.009
10	7.5e-5±6.2e-6 (6.1e-4±1e-4)	0.015±0.001 (0.050±0.005)	0.134±0.036 (168.696±7.896)	0.490±0.015
20	6.8e-5±8.3e-6 (5.5e-4±4.6e-5)	0.011±0.001 (0.034±0.002)	0.149±0.039 (162.488±37.734)	0.668±0.037
30	6.6e-5±5.5e-6 (9.2e-4±1.9e-4)	0.015±0.002 (0.046±0.008)	0.124±0.033 (166.242±45.394)	0.695±0.008
40	6.6e-5±5.5e-6 (9.2e-4±1.9e-4)	0.015±0.002 (0.046±0.008)	0.125±0.040 (136.559±3.089)	0.682±0.011
50	6.6e-5±5.5e-6 (9.2e-4±1.9e-4)	0.015±0.002 (0.046±0.008)	0.125±0.033 (176.246±49.894)	0.691±0.019
<b>Convex QCQP: <math>n = 100, n_{eq} = 50, n_{ineq} = 50</math></b>				
0	7.2e-5±4.5e-7 (7.9e-4±9.7e-5)	37.174±0.148 (45.584±0.395)	0.648±0.003 (284.443±35.772)	0.461±0.011
10	6.2e-5±2.8e-6 (8.1e-4±6.6e-5)	0.032±0.002 (0.405±0.062)	0.475±0.008 (193.722±28.776)	0.536±0.037
20	6.2e-5±8.0e-6 (1.1e-03±1.8e-4)	0.032±0.002 (0.411±0.052)	0.451±0.002 (223.545±27.455)	0.771±0.047
30	6.2e-5±5.8e-6 (2.1e-03±5.7e-4)	0.035±0.003 (0.432±0.070)	0.457±0.009 (213.981±18.431)	0.757±0.028
40	6.2e-5±5.8e-6 (2.1e-03±5.7e-4)	0.035±0.003 (0.432±0.070)	0.440±0.006 (200.018±28.512)	0.729±0.015
50	6.2e-5±5.8e-6 (2.1e-03±5.7e-4)	0.035±0.003 (0.432±0.070)	0.446±0.003 (174.800±28.764)	0.765±0.026
<b>Convex SOCP: <math>n = 100, n_{eq} = 50, n_{ineq} = 50</math></b>				
0	4.8e-5±2e-5 (4.7e-4±7.4e-5)	181.38±0.092 (194.957±0.896)	0.416±0.016 (354.401±86.816)	0.406±0.032
10	6.6e-5±7.3e-6 (6.2e-4±6.3e-5)	0.136±0.010 (1.006±0.143)	0.288±0.008 (252.299±56.041)	0.532±0.050
20	6.2e-5±2.2e-6 (5.6e-4±2.2e-5)	0.186±0.020 (1.914±0.439)	0.289±0.003 (187.842±26.580)	0.717±0.037
30	6.3e-5±6.1e-6 (6.7e-4±9.9e-5)	0.159±0.005 (1.889±0.376)	0.336±0.036 (272.440±52.514)	0.780±0.024
40	6.3e-5±6.1e-6 (6.7e-4±9.9e-5)	0.159±0.005 (1.889±0.376)	0.297±0.010 (214.608±17.089)	0.804±0.015
50	6.3e-5±6.1e-6 (6.7e-4±9.9e-5)	0.159±0.005 (1.889±0.376)	0.297±0.008 (259.888±30.394)	0.771±0.043

Table B.4: Numerical results on 2000 instances of smooth convex problems with varying numbers of tracked iterations.

## 534 B.5 FSNet with different values of $\rho$

535 This section investigates the impact of the value of penalty weight  $\rho$  on the performance of FSNet  
536 (see Section 3.1). Table B.5 shows that the constraint violations are near-zero for all choices of  $\rho$ ,  
537 since the feasibility-seeking step always tries to enforce the solution’s feasibility regardless of the  
538 NN’s prediction. This is an advantage of FSNet, as the feasibility is guaranteed by a well-designed  
539 feasibility-seeking step, rather than being left to the NN alone. While the optimality gaps are  
540 similar across different  $\rho$ , both training and inference times decrease as  $\rho$  increases (Figure B.7).  
541 Actually, a larger penalty weight strongly penalizes the distance between the solutions before and  
542 after the feasibility-seeking step, driving the NN to produce outputs that are closer to the feasible  
543 set. This reduces the number of iterations required by the feasibility-seeking step, leading to lower  
544 computational time for the entire process. For example, setting  $\rho = 50$  makes the training  $1.4\text{--}1.6\times$   
545 faster and inference  $1.3\text{--}1.7\times$  faster compared to  $\rho = 0$ . However, excessively large values of  $\rho$  may  
546 cause the optimization to overemphasize feasibility at the expense of optimality, potentially slowing  
547 convergence of the network parameters.

$\rho$	Total vio. Mean (Max)	Optimality Gap (%) Mean (Max)	Runtime (s) Batch (Sequential)	Training time (h)
<b>Convex QP: <math>n = 100, n_{eq} = 100, n_{ineq} = 100</math></b>				
0	$5.8\text{e-}05 \pm 6.2\text{e-}06$ ( $5.8\text{e-}04 \pm 1.0\text{e-}04$ )	$0.008 \pm 0.001$ ( $0.029 \pm 0.003$ )	$0.183 \pm 0.055$ ( $261.704 \pm 13.474$ )	$1.075 \pm 0.027$
0.5	$6.5\text{e-}05 \pm 1.0\text{e-}05$ ( $9.5\text{e-}04 \pm 2.4\text{e-}04$ )	$0.009 \pm 0.001$ ( $0.031 \pm 0.004$ )	$0.150 \pm 0.034$ ( $236.686 \pm 50.218$ )	$0.827 \pm 0.076$
2.5	$6.6\text{e-}05 \pm 8.6\text{e-}06$ ( $7.8\text{e-}04 \pm 1.4\text{e-}04$ )	$0.011 \pm 0.002$ ( $0.041 \pm 0.004$ )	$0.183 \pm 0.014$ ( $196.120 \pm 47.639$ )	$0.747 \pm 0.044$
5.0	$6.6\text{e-}05 \pm 5.5\text{e-}06$ ( $9.2\text{e-}04 \pm 1.9\text{e-}04$ )	$0.015 \pm 0.002$ ( $0.046 \pm 0.008$ )	$0.142 \pm 0.027$ ( $186.968 \pm 34.144$ )	$0.687 \pm 0.020$
10.0	$7.1\text{e-}05 \pm 6.9\text{e-}06$ ( $9.6\text{e-}04 \pm 2.0\text{e-}04$ )	$0.014 \pm 0.001$ ( $0.042 \pm 0.004$ )	$0.117 \pm 0.029$ ( $152.014 \pm 22.923$ )	$0.676 \pm 0.022$
20.0	$6.7\text{e-}05 \pm 2.4\text{e-}06$ ( $1.2\text{e-}03 \pm 2.5\text{e-}04$ )	$0.011 \pm 0.002$ ( $0.042 \pm 0.003$ )	$0.133 \pm 0.050$ ( $165.496 \pm 23.169$ )	$0.649 \pm 0.013$
50.0	$5.9\text{e-}05 \pm 1.1\text{e-}05$ ( $9.3\text{e-}04 \pm 2.5\text{e-}04$ )	$0.018 \pm 0.006$ ( $0.074 \pm 0.016$ )	$0.104 \pm 0.013$ ( $156.569 \pm 53.576$ )	$0.661 \pm 0.012$
<b>Convex QCQP: <math>n = 100, n_{eq} = 50, n_{ineq} = 50</math></b>				
0	$7.2\text{e-}05 \pm 2.7\text{e-}06$ ( $7.3\text{e-}04 \pm 1.6\text{e-}04$ )	$0.023 \pm 0.002$ ( $0.315 \pm 0.064$ )	$0.576 \pm 0.004$ ( $257.309 \pm 15.105$ )	$1.039 \pm 0.037$
0.5	$5.7\text{e-}05 \pm 1.0\text{e-}05$ ( $1.0\text{e-}03 \pm 1.5\text{e-}04$ )	$0.028 \pm 0.002$ ( $0.412 \pm 0.075$ )	$0.487 \pm 0.015$ ( $216.307 \pm 32.152$ )	$0.811 \pm 0.010$
2.5	$7.1\text{e-}05 \pm 7.8\text{e-}06$ ( $1.6\text{e-}03 \pm 3.6\text{e-}04$ )	$0.030 \pm 0.001$ ( $0.411 \pm 0.073$ )	$0.464 \pm 0.004$ ( $187.807 \pm 11.011$ )	$0.771 \pm 0.024$
5.0	$6.2\text{e-}05 \pm 5.8\text{e-}06$ ( $2.1\text{e-}03 \pm 5.7\text{e-}04$ )	$0.035 \pm 0.003$ ( $0.432 \pm 0.070$ )	$0.451 \pm 0.008$ ( $180.282 \pm 32.059$ )	$0.743 \pm 0.014$
10.0	$6.1\text{e-}05 \pm 6.6\text{e-}06$ ( $1.5\text{e-}03 \pm 1.2\text{e-}04$ )	$0.038 \pm 0.003$ ( $0.417 \pm 0.058$ )	$0.450 \pm 0.014$ ( $222.192 \pm 56.348$ )	$0.715 \pm 0.006$
20.0	$5.8\text{e-}05 \pm 6.8\text{e-}06$ ( $2.4\text{e-}03 \pm 6.9\text{e-}04$ )	$0.041 \pm 0.007$ ( $0.466 \pm 0.081$ )	$0.437 \pm 0.009$ ( $240.089 \pm 33.913$ )	$0.762 \pm 0.052$
50.0	$5.3\text{e-}05 \pm 1.5\text{e-}06$ ( $2.6\text{e-}03 \pm 7.5\text{e-}04$ )	$0.049 \pm 0.016$ ( $0.486 \pm 0.088$ )	$0.447 \pm 0.015$ ( $225.130 \pm 36.618$ )	$0.716 \pm 0.031$
<b>Convex SOCP: <math>n = 100, n_{eq} = 50, n_{ineq} = 50</math></b>				
0	$5.3\text{e-}05 \pm 8.2\text{e-}06$ ( $7.1\text{e-}04 \pm 1.2\text{e-}04$ )	$0.356 \pm 0.097$ ( $1.490 \pm 0.278$ )	$0.360 \pm 0.012$ ( $310.059 \pm 71.106$ )	$0.976 \pm 0.044$
0.5	$7.2\text{e-}05 \pm 8.3\text{e-}06$ ( $6.6\text{e-}04 \pm 7.6\text{e-}05$ )	$0.180 \pm 0.012$ ( $2.347 \pm 0.514$ )	$0.325 \pm 0.017$ ( $237.029 \pm 72.636$ )	$0.921 \pm 0.014$
2.5	$6.6\text{e-}05 \pm 9.2\text{e-}06$ ( $7.0\text{e-}04 \pm 1.1\text{e-}04$ )	$0.174 \pm 0.019$ ( $0.978 \pm 0.109$ )	$0.294 \pm 0.007$ ( $206.250 \pm 15.332$ )	$0.797 \pm 0.028$
5.0	$6.3\text{e-}05 \pm 6.1\text{e-}06$ ( $6.7\text{e-}04 \pm 9.9\text{e-}05$ )	$0.159 \pm 0.005$ ( $1.889 \pm 0.376$ )	$0.307 \pm 0.014$ ( $281.764 \pm 18.464$ )	$0.837 \pm 0.005$
10.0	$7.9\text{e-}05 \pm 7.0\text{e-}07$ ( $7.7\text{e-}04 \pm 8.2\text{e-}05$ )	$0.174 \pm 0.020$ ( $1.365 \pm 0.251$ )	$0.291 \pm 0.011$ ( $198.642 \pm 15.576$ )	$0.725 \pm 0.011$
20.0	$6.1\text{e-}05 \pm 4.1\text{e-}06$ ( $9.5\text{e-}04 \pm 1.2\text{e-}04$ )	$0.184 \pm 0.017$ ( $0.994 \pm 0.094$ )	$0.287 \pm 0.013$ ( $197.955 \pm 33.984$ )	$0.714 \pm 0.041$
50.0	$6.3\text{e-}05 \pm 9.1\text{e-}06$ ( $7.4\text{e-}04 \pm 9.2\text{e-}05$ )	$0.190 \pm 0.016$ ( $1.093 \pm 0.192$ )	$0.269 \pm 0.015$ ( $172.091 \pm 41.039$ )	$0.694 \pm 0.027$

Table B.5: Numerical results of FSNet on 2000 instances of smooth convex problems with varying  $\rho$ .

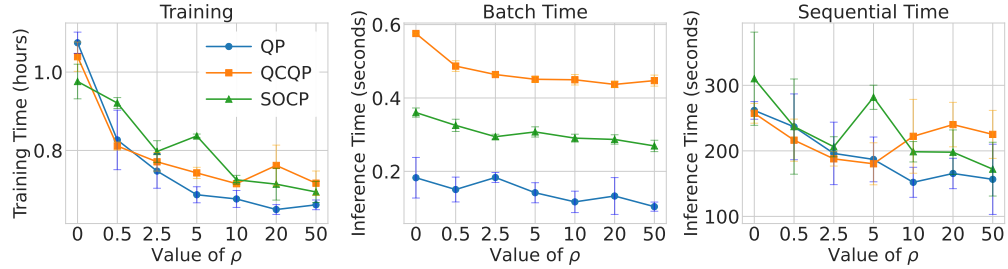


Figure B.7: Training time and batch inference time of FSNet with different values of  $\rho$ .

## 548 B.6 More experiments for DC3

549 In [13], the authors mention that unrolling more correction steps may improve the feasibility of the  
 550 prediction. To evaluate this, we ran DC3 with different numbers of correction steps on 2000 test  
 551 instances of smooth convex SOCP. As shown in Table B.6, increasing the number of correction steps  
 552 only yields a marginal reduction in inequality violations (due to the default small step size of  $10^{-7}$ ),  
 553 while significantly increasing runtime.

554 Next, we evaluate the performance of DC3 with different correction step sizes. We fixed the number  
 555 of maximum correction steps and changed the stepsize to obtain the results in Table B.7. Stepsizes  
 556 up to  $10^{-5}$  achieve limited violation reduction, whereas stepsizes of  $10^{-4}$  and  $10^{-3}$  substantially  
 557 mitigate violations. However, overly big stepsize (e.g.,  $10^{-2}$ ) causes the correction procedure to  
 558 diverge. Note that all experiments here are in inference time. During training, a large correction  
 559 stepsize can destabilize the optimization, as noted in [13]. This demonstrates the DC3’s sensitivity  
 560 to the correction stepsize, whereas our feasibility-seeking step, implemented via L-BGFS with a  
 561 backtracking line search, mitigates such sensitivity and delivers robust performance across diverse  
 562 problems using the same hyperparameters.

Max Correction Steps	Eq. Violation	Ineq. Violation	Optimality Gap (%)	Time (s)
10	2.30e-14	0.024	0.051	0.095
50	2.30e-14	0.024	0.051	0.474
100	2.27e-14	0.024	0.052	0.948
200	2.28e-14	0.023	0.052	1.896
500	2.28e-14	0.022	0.054	4.738
1000	2.34e-14	0.020	0.057	9.526
2000	2.28e-14	0.016	0.061	19.046
5000	2.27e-14	0.010	0.070	48.024

Table B.6: Sensitivity of DC3 performance to the maximum number of correction steps on 2000 instances of smooth convex SOCP problem, with fixed correction stepsize of  $10^{-7}$ .

Correction Stepsize	Eq. Violation	Ineq. Violation	Objective Gap (%)	Time (s)
1e-07	2.30e-14	0.024	0.051	0.338
1e-06	2.30e-14	0.024	0.052	0.121
1e-05	2.28e-14	0.020	0.056	0.095
1e-04	2.28e-14	0.004	0.081	0.096
1e-03	2.31e-14	0.000	0.130	0.095
1e-02	-3.52e+12	2.89e+30	1e58	0.095

Table B.7: Sensitivity of DC3 performance to the correction stepsize on 2000 instances of smooth convex SOCP problem, with fixed maximum correction steps of 10.



## 563 C Hyperparameters

564 For DC3 and Penalty methods, we adopt the hyperparameters as tuned in [13]. For the Adaptive  
 565 Penalty method, we performed a brief manual sweep to identify settings yielding strong empirical  
 566 performance. For FSNet, we configured the L-BFGS solver with a sufficiently large maximum-  
 567 iteration budget and memory size to guarantee convergence on all the problem classes. We also  
 568 evaluated several neural network sizes and learning rates to ensure sufficient expressive capacity and  
 569 stable optimization for all problem classes. Table C.1 lists all hyperparameter values.

Hyperparameter	Value
<i>Shared:</i>	
Learning rate	5e-4
Learning rate decay	0.5
Number of hidden layers	4
Nonlinearity	SiLU
Number of hidden neurons per layer	1024
Train/validation/test ratio	0.7/0.1/0.2
Minibatch size	512
Random seeds	2025, 2027, 2029
<i>FS:</i>	
Equality violation weight	10
Inequality violation weight	10
Max L-BFGS iterations	50
L-BFGS memory	30
Epoch	100
Learning rate decay steps	2000
<i>Penalty:</i>	
Equality violation weight	50
Inequality violation weight	50
Epoch	1000
Learning rate decay steps	4000
<i>Adaptive Penalty:</i>	
Initial equality violation weight	30
Initial inequality violation weight	30
Max equality violation weight	500
Max inequality violation weight	500
Increasing rate	2
Epoch	1000
Learning rate decay steps	4000
<i>DC3:</i>	
Equality violation weight	10
Inequality violation weight	10
Correction stepsize	1e-7
Max correction stepsize	10
Correction momentum	0.5
Epoch	1000
Learning rate decay steps	4000

Table C.1: Hyperparameters

## D Analysis and Proofs

### D.1 Proof of Theorem 2

**Theorem 2.** Suppose that  $\mathbb{E}_t [\|\tilde{\nabla} L(\theta_t)\|_2^2] \leq G$  for all  $t$  and that the feasibility-seeking step is unrolled for  $K$  steps in every forward pass. Let  $\gamma$  be as defined in Theorem 1. Under Assumption 1, 2, and 3, after running the SGD for  $T$  iterations with step size  $\eta = \eta_0/\sqrt{T}$  for some  $\eta_0 > 0$ , there is at least one iteration  $t \in \{0, \dots, T-1\}$  satisfying

$$\mathbb{E} [\|\nabla L(\theta_t)\|_2^2] \leq \mathcal{O}(T^{-1/2}), \quad (11)$$

$$\mathbb{E} [\|\nabla \mathcal{L}(\theta_t)\|_2] \leq \mathcal{O}(T^{-1/4} + \gamma^K). \quad (12)$$

**Proof. Gradient bound of finite-unrolling loss.** For the first part of the theorem, we follow the analysis of the SGD for a  $L$ -smooth function. By the quadratic upper bound of the  $L$ -smooth function, we have

$$L(\theta_{t+1}) \leq L(\theta_t) + \langle \nabla L(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{L_N}{2} \|\theta_{t+1} - \theta_t\|_2^2.$$

Observing from the SGD update (10) that  $\theta_{t+1} - \theta_t = -\eta \tilde{\nabla} L(\theta_t)$ , taking the conditional expectation given all randomness up to time  $t$ , and rearranging terms, we obtain

$$\|\nabla L(\theta_t)\|_2^2 \leq \frac{1}{\eta} (L(\theta_t) - \mathbb{E}_t[L(\theta_{t+1})]) + \frac{L_N \eta}{2} \mathbb{E}_t [\|\tilde{\nabla} L(\theta_t)\|_2^2].$$

Taking the full expectation for both sides and using the assumption  $\mathbb{E}_t [\|\tilde{\nabla} L(\theta_t)\|_2^2] \leq G$  yields

$$\mathbb{E} [\|\nabla L(\theta_t)\|_2^2] \leq \frac{1}{\eta} (L(\theta_t) - \mathbb{E}[L(\theta_{t+1})]) + \frac{L_N \eta G}{2}.$$

Finally, we take the average over  $t \in 0, \dots, T-1$  and use the telescoping sum to get

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla L(\theta_t)\|_2^2] \leq \frac{L(\theta_0) - L^*}{\eta T} + \frac{L_N \eta G}{2}.$$

Since  $\min_{t \in \{0, \dots, T-1\}} \|\nabla L(\theta_t)\|_2^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla L(\theta_t)\|_2^2$ , we obtain the first statement of the theorem by replacing the stepsize  $\eta = \eta_0/\sqrt{T}$ .

**Gradient bound of infinite-unrolling loss.** For any  $\theta$ , we consider the norm of the gradient difference:

$$\begin{aligned} \|\nabla \mathcal{L}(\theta) - \nabla L(\theta)\|_2 &= \left\| \frac{1}{S} \sum_{i=1}^S \nabla_{\theta} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)})) - \nabla_{\theta} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}^K(x^{(i)})) \right\|_2 \\ &\leq \frac{1}{S} \sum_{i=1}^S \left\| \nabla_{\theta} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)})) - \nabla_{\theta} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}^K(x^{(i)})) \right\|_2, \quad (\text{D.1}) \end{aligned}$$

which follows from the triangle inequality. We consider one element of the sum, and let  $g := \|\nabla_{\theta} F(y_{\theta}(x), \hat{y}_{\theta}(x)) - \nabla_{\theta} F(y_{\theta}(x), \hat{y}_{\theta}^K(x))\|$ . Then, via the chain rule, we have

$$\begin{aligned} g &= \|(J_{\theta} y_{\theta}(x))^{\top} \nabla_y F(y_{\theta}(x), \hat{y}_{\theta}(x)) + (J_{\theta} y_{\theta}(x))^{\top} (J_y \text{FS}(y_{\theta}(x); x))^{\top} \nabla_{\hat{y}} F(y_{\theta}(x), \hat{y}_{\theta}(x)) \\ &\quad - (J_{\theta} y_{\theta}(x))^{\top} \nabla_y F(y_{\theta}(x), \hat{y}_{\theta}^K(x)) - (J_{\theta} y_{\theta}(x))^{\top} (J_y \text{FS}^K(y_{\theta}(x); x))^{\top} \nabla_{\hat{y}} F(y_{\theta}(x), \hat{y}_{\theta}^K(x))\|_2 \\ &\leq B_N \|\nabla_y F(y_{\theta}(x), \hat{y}_{\theta}(x)) - \nabla_y F(y_{\theta}(x), \hat{y}_{\theta}^K(x))\|_2 \\ &\quad + B_N \|(J_y \text{FS}(y_{\theta}(x); x))^{\top} \nabla_{\hat{y}} F(y_{\theta}(x), \hat{y}_{\theta}(x)) - (J_y \text{FS}^K(y_{\theta}(x); x))^{\top} \nabla_{\hat{y}} F(y_{\theta}(x), \hat{y}_{\theta}^K(x))\|_2. \end{aligned}$$

where the inequality follows from the bounded Jacobian in Assumption 2.

590 The first term becomes (by expansion of the gradient terms):

$$\|\nabla_y F(y_\theta(x), \hat{y}_\theta(x)) - \nabla_y F(y_\theta(x), \hat{y}_\theta^K(x))\|_2 = \rho \|\hat{y}_\theta(x) - \hat{y}_\theta^K(x)\|_2.$$

591 The second term:

$$\begin{aligned} & \|(J_y \text{FS}(y_\theta(x); x))^\top \nabla_{\hat{y}} F(y_\theta(x), \hat{y}_\theta(x)) - (J_y \text{FS}^K(y_\theta(x); x))^\top \nabla_{\hat{y}} F(y_\theta(x), \hat{y}_\theta^K(x))\|_2 \\ &= \|(J_y \text{FS}(y_\theta(x); x))^\top \nabla_{\hat{y}} F(y_\theta(x), \hat{y}_\theta(x)) - (J_y \text{FS}^K(y_\theta(x); x))^\top \nabla_{\hat{y}} F(y_\theta(x), \hat{y}_\theta(x)) \\ & \quad + (J_y \text{FS}^K(y_\theta(x); x))^\top \nabla_{\hat{y}} F(y_\theta(x), \hat{y}_\theta(x)) - (J_y \text{FS}^K(y_\theta(x); x))^\top \nabla_{\hat{y}} F(y_\theta(x), \hat{y}_\theta^K(x))\|_2 \\ &\leq \|J_y \text{FS}(y_\theta(x); x) - J_y \text{FS}^K(y_\theta(x); x)\|_2 \|\nabla_{\hat{y}} F(y_\theta(x), \hat{y}_\theta(x))\|_2 \\ & \quad + \|J_y \text{FS}^K(y_\theta(x); x)\|_2 \|\nabla_{\hat{y}} F(y_\theta(x), \hat{y}_\theta(x)) - \nabla_{\hat{y}} F(y_\theta(x), \hat{y}_\theta^K(x))\|_2 \\ &\leq B_F L_{FS} \|\hat{y}_\theta(x) - \hat{y}_\theta^K(x)\|_2 + B_{FS} L_F \|\hat{y}_\theta(x) - \hat{y}_\theta^K(x)\|_2 \\ &= (B_F L_{FS} + B_{FS} L_F) \|\hat{y}_\theta(x) - \hat{y}_\theta^K(x)\|_2, \end{aligned}$$

592 where the second inequality follows from the smoothness of the feasibility-seeking mapping and  
593 function  $F(y_\theta(x), \hat{y}_\theta(x))$  (Assumption 2).

594 Let  $\nu = \rho + B_F L_{FS} + B_{FS} L_F$ . Thus,  $g$  is upper bounded by

$$g \leq \nu \|\hat{y}_\theta^K(x) - \hat{y}_\theta(x)\|_2.$$

595 Using this upper bound and Theorem 1, we can rewrite the gradient difference (D.1) as follows:

$$\begin{aligned} \|\nabla \mathcal{L}(\theta) - \nabla L(\theta)\|_2 &\leq \frac{1}{S} \sum_{i=1}^S \nu \|\hat{y}_\theta^K(x^{(i)}) - \hat{y}_\theta(x^{(i)})\|_2 \\ &\leq \frac{1}{S} \sum_{i=1}^S \frac{8\nu\eta_\phi L_\phi^2}{\mu_\phi^2} (\phi(y_\theta^{(i)}; x^{(i)}) - \phi(\hat{y}_\theta(x^{(i)}); x^{(i)})), \end{aligned} \quad (\text{D.2})$$

596 where  $\hat{y}_\theta(x^{(i)}) = \text{FS}(y_\theta(x^{(i)}); x^{(i)})$ .

597 Let  $\Phi(\theta) := \frac{1}{S} \sum_{i=1}^S \frac{8\nu\eta_\phi L_\phi^2}{\mu_\phi^2} \gamma^{K-1} (\phi(y_\theta(x^{(i)}); x^{(i)}) - \phi(\hat{y}_\theta(x^{(i)}); x^{(i)}))$ . Using the inverse triangle  
598 inequality, the inequality (D.2) leads to

$$\|\nabla \mathcal{L}(\theta)\|_2 \leq \|\nabla L(\theta)\|_2 + \Phi(\theta) \gamma^{K-1}.$$

599 We consider the time  $t$  and use (11) to obtain

$$\begin{aligned} \mathbb{E}[\|\nabla \mathcal{L}(\theta_t)\|_2] &\leq \mathbb{E}[\|\nabla L(\theta_t)\|_2] + \mathbb{E}[\Phi(\theta_t)] \gamma^{K-1} \\ &\leq \mathcal{O}(T^{-1/4}) + \mathbb{E}[\Phi(\theta_t)] \gamma^{K-1}. \end{aligned} \quad (\text{D.3})$$

600 Finally, we need to show that  $\mathbb{E}[\Phi(\theta_t)]$  is upper bounded. We have

$$\mathbb{E}[\Phi(\theta_t)] = \frac{1}{S} \sum_{i=1}^S \frac{8\nu\eta_\phi L_\phi^2}{\mu_\phi^2} \mathbb{E} \left[ \phi(y_\theta(x^{(i)}); x^{(i)}) - \phi(\hat{y}_\theta(x^{(i)}); x^{(i)}) \right]. \quad (\text{D.4})$$

601 Since Theorem 1 shows that the gradient descent converges to the minimum, we have  
602  $\phi(y_\theta(x^{(i)}); x^{(i)}) \geq \phi(\hat{y}_\theta(x^{(i)}); x^{(i)})$ . Thus, it suffices to show  $\mathbb{E}[\phi(y_\theta(x^{(i)}); x^{(i)})]$  is upper bounded  
603 for any  $i$ . The argument proceeds as follows: We first show that the NN parameters remain bounded  
604 during training with SGD, which implies that the network output  $y_\theta(x^{(i)})$  is also bounded. From  
605 there, we conclude that the expectation  $\mathbb{E}[\phi(y_\theta(x^{(i)}); x^{(i)})]$  is bounded from above.

606 *Bound of the NN parameters:*

607 We first establish the bounds of the NN parameters during training with SGD, which via (10) exhibits

$$\|\theta_{t+1} - \theta_t\|_2 = \eta \|\tilde{\nabla} L(\theta_t)\|_2, \quad t = 0, \dots, T-1.$$

608 Taking the sum over  $j = 0, \dots, t-1$  and using triangle inequality and the telescoping sum, we have

$$\|\theta_t - \theta_0\|_2 \leq \sum_{j=0}^{t-1} \|\theta_{j+1} - \theta_j\|_2 = \sum_{j=0}^{t-1} \eta \|\tilde{\nabla} L(\theta_j)\|_2.$$

609 Taking the square for both sides and using the Cauchy-Schwarz inequality yields

$$\|\theta_t - \theta_0\|_2^2 \leq \sum_{j=0}^{t-1} \eta^2 \sum_{j=0}^{t-1} \|\tilde{\nabla} L(\theta_j)\|_2^2 \leq \eta^2 T \sum_{j=0}^{t-1} \|\tilde{\nabla} L(\theta_j)\|_2^2.$$

610 Using the assumption of bounded variance of the stochastic gradients, we obtain

$$\mathbb{E}_t[\|\theta_t - \theta_0\|_2^2] \leq \eta^2 T^2 G.$$

611 By replacing  $\eta = \eta_0/\sqrt{T}$  and taking a regular expectation, we finally obtain

$$\mathbb{E}[\|\theta_t - \theta_0\|_2^2] \leq \eta_0^2 T G. \quad (\text{D.5})$$

612 *Bound of the NN output:*

613 For notational brevity, we denote  $y_0 = y_{\theta_0}(x)$  and  $y_t = y_{\theta_t}(x)$ , which are outputs of NNs with  
 614 initialized parameter  $\theta_0$  and parameter  $\theta_t$  at time  $t$ . By the mean value theorem and the bounded  
 615 Jacobian  $\|J_{\theta} y_{\theta}(x)\|_2 \leq B_N$  in Assumption 2, we have

$$\|y_t - y_0\|_2 \leq B_N \|\theta_t - \theta_0\|_2,$$

616 which, squaring both sides, leads to

$$\mathbb{E}[\|y_t - y_0\|_2^2] \leq B_N^2 \mathbb{E}[\|\theta_t - \theta_0\|_2^2] \leq B_N^2 \eta_0^2 T G. \quad (\text{D.6})$$

617 *Bound of  $\mathbb{E}[\phi(y_t; x)]$ :*

618 By the  $L$ -smoothness of  $\phi(y_t; x)$ , we have the quadratic upper bound:

$$\phi(y_t; x) \leq \phi(y_0; x) + \langle \nabla_y \phi(y_0; x), y_t - y_0 \rangle + \frac{L_{\phi}}{2} \|y_t - y_0\|_2^2.$$

619 Taking the expectation on both sides and using (D.6) yields

$$\mathbb{E}[\phi(y_t; x)] \leq \phi(y_0; x) + B_N \eta \sqrt{T G} \|\nabla_y \phi(y_0; x)\|_2 + \frac{L_{\phi}}{2} B_N^2 \eta_0^2 T G < \infty.$$

620 Therefore, we have that  $\mathbb{E}[\phi(y_{\theta}(x^{(i)}); x^{(i)})]$  is upper bounded for any  $i$ , which implies that  
 621  $\mathbb{E}[\phi(\hat{y}_{\theta}(x^{(i)}); x^{(i)})]$  is similarly bounded given that  $\phi(\hat{y}_{\theta}(x^{(i)}); x^{(i)}) \leq \phi(y_{\theta}(x^{(i)}); x^{(i)})$  by The-  
 622 orem 1 Finally, by (D.4), there exists a constant  $\kappa > 0$  such that  $\mathbb{E}[\Phi(\theta_t)] \leq \kappa$  for  $t = 0, \dots, T-1$ .  
 623 Using this in (D.3), we complete the second statement of the theorem.  $\square$

#### 624 D.1.1 Proof of Lemma 1

625 **Lemma 1.** Suppose SGD converges to  $\theta^*$ , where  $J(\theta^*) := ((J_{\theta} y_{\theta^*}(x^{(1)}))^{\top} \dots (J_{\theta} y_{\theta^*}(x^{(S)}))^{\top})$   
 626 has full column rank. Then, the prediction of the NN  $y_{\theta}(x^{(i)})$  is the stationary point of  
 627  $F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)}))$  with  $\hat{y}_{\theta}(x^{(i)}) = \text{FS}(y_{\theta}(x^{(i)}); x^{(i)})$  for all  $i \in \{1, \dots, S\}$ .

628 *Proof.* At the stationary parameter  $\theta^*$ , the gradient of the loss function vanishes:

$$\begin{aligned} 0 = \nabla \mathcal{L}(\theta^*) &= \frac{1}{S} \sum_{i=1}^S (J_{\theta} y_{\theta^*}(x^{(i)}))^{\top} \left( \nabla_y F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)})) \right. \\ &\quad \left. + (J_y \text{FS}(y_{\theta}(x^{(i)}); x^{(i)}))^{\top} \nabla_{\hat{y}} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)})) \right). \end{aligned}$$

629 Let  $v^{(i)} = \nabla_y F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)})) + (J_y \text{FS}(y^{(i)}; x^{(i)}))^{\top} \nabla_{\hat{y}} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)}))$ . The stationarity  
 630 results in

$$J(\theta^*) \begin{pmatrix} v^{(1)} \\ \vdots \\ v^{(S)} \end{pmatrix} = 0.$$

631 Since  $J(\theta^*)$  has full column rank, this implies  $(v^{(1)}, \dots, v^{(S)})^{\top} = 0$ . Thus, the NN output  $y_{\theta^*}(x^{(i)})$   
 632 and  $\hat{y}_{\theta}(x^{(i)}) = \text{FS}(y_{\theta}(x^{(i)}); x^{(i)})$  satisfy

$$v^{(i)} = \nabla_y F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)})) + (J_y \text{FS}(y^{(i)}; x^{(i)}))^{\top} \nabla_{\hat{y}} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)})) = 0,$$

633 which is the stationary condition of  $F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}(x^{(i)}))$ .  $\square$

### 634 D.1.2 Proof of Theorem 3

635 **Theorem 3.** *Given specific parameters  $x$ , suppose that the NN predicts  $y_\tau$  which is a global minimizer*  
 636 *of  $F(y, \hat{y}; \rho_\tau) = f(\hat{y}; x) + \frac{\rho_\tau}{2} \|y - \hat{y}\|_2^2$  with  $\hat{y} = \text{FS}(y; x)$ ; that the mapping FS is continuous; and*  
 637 *that  $0 < \rho_\tau < \rho_{\tau+1}, \forall \tau$  and  $\rho_\tau \rightarrow \infty$ . Then, every limit point  $(y^*, \hat{y}^*)$  of the sequence  $\{(y_\tau, \hat{y}_\tau)\}$*   
 638 *admits  $y^* = \hat{y}^*$ , and  $\hat{y}^*$  is a global minimizer of the optimization problem (1).*

639 *Proof.* Let  $\bar{y}$  be a global solution to problem (1), that is

$$f(\bar{y}; x) \leq f(\hat{y}; x)$$

640 for all the feasible points  $\hat{y}$ . We also have that  $\bar{y} = \text{FS}(\bar{y}; x)$  given that  $\bar{y}$  is already feasible. By  
 641 the global optimality of  $(y_\tau, \hat{y}_\tau)$ , we have that  $F(y_\tau, \hat{y}_\tau; \rho_\tau) \leq F(\bar{y}, \bar{y}; \rho_\tau)$  which results in the  
 642 inequality

$$f(\hat{y}_\tau; x) + \frac{\rho_\tau}{2} \|y_\tau - \hat{y}_\tau\|_2^2 \leq f(\bar{y}; x) + \frac{\rho_\tau}{2} \|\bar{y} - \bar{y}\|_2^2 = f(\bar{y}; x). \quad (\text{D.7})$$

643 By rearranging this inequality, we obtain

$$\|y_\tau - \hat{y}_\tau\|_2^2 \leq \frac{2}{\rho_\tau} (f(\bar{y}; x) - f(\hat{y}_\tau; x)). \quad (\text{D.8})$$

644 Suppose that  $y^*$  is a limit point of  $\{y_\tau\}_{\tau \geq 0}$ , so that there is an infinite subsequence  $\mathcal{T}$  such that

$$\lim_{\tau \in \mathcal{T}} y_\tau = y^*.$$

645 As the feasibility seeking mapping is continuous by assumption, we can define  $\hat{y}^* :=$   
 646  $\lim_{\tau \in \mathcal{T}} \text{FS}(y_\tau; x)$ . By the continuity, we have

$$\lim_{\tau \in \mathcal{T}} \|y_\tau - \hat{y}_\tau\|_2^2 = \lim_{\tau \in \mathcal{T}} \|y_\tau - \text{FS}(y_\tau; x)\|_2^2 = \|y^* - \hat{y}^*\|_2^2.$$

647 Taking the limit as  $\tau \rightarrow \infty, \tau \in \mathcal{T}$  on both sides of (D.8) yields

$$\|y^* - \hat{y}^*\|_2^2 = \lim_{\tau \in \mathcal{T}} \|y_\tau - \hat{y}_\tau\|_2^2 \leq \lim_{\tau \in \mathcal{T}} \frac{2}{\rho_\tau} (f(\bar{y}; x) - f(\hat{y}_\tau; x)) = 0.$$

648 Therefore, we have that  $y^* = \hat{y}^*$ . Note that  $\hat{y}^*$  is the output of the feasibility-seeking mapping, so  $y^*$   
 649 and  $\hat{y}^*$  are feasible. Moreover, by taking the limit as  $\tau \rightarrow \infty, \tau \in \mathcal{T}$  in (D.7), we have

$$f(\hat{y}^*) \leq f(\hat{y}^*) + \lim_{\tau \in \mathcal{T}} \frac{\rho_\tau}{2} \|y_\tau - \hat{y}_\tau\|_2^2 \leq f(\bar{y}; x).$$

650 Since  $\hat{y}^*$  is a feasible point with an objective value no larger than that of the global solution  $\bar{y}$ , it must  
 651 itself be a global solution.  $\square$

652 Using a similar argument, we can show that  $\hat{y}^*$  is a minimizer of Problem 1 under the same assump-  
 653 tions, even when  $\rho = 0$ . However, in practice, we often choose  $\rho > 0$  for the reasons discussed in  
 654 Section 3.1 and Appendix B.

## 655 D.2 Analysis of Truncated Unrolled Differentiation

656 In this section, we build intuition for why truncating backpropagation remains effective in practice.  
 657 Specifically, if the violation function  $\phi(s, x)$  is locally strongly convex, the bias introduced by  
 658 truncation decays exponentially with the truncation depth  $K'$ . Consequently, a relatively small  $K'$   
 659 suffices to yield a high-quality gradient estimate.

660 **Assumption 4.** *The violation function  $s \mapsto \phi(s; x)$  is twice differentiable. For  $s_K \in \mathbb{R}^n$ , there exists*  
 661 *a neighborhood  $\mathcal{N}$  of  $s_K$  where  $s \mapsto \phi(s; x)$  is  $\mu'$ -strongly convex. Moreover, for some  $K_0 \in \mathbb{N}$ , the*  
 662 *iterates satisfy  $s_k \in \mathcal{N}$  for  $k \geq K_0$ .*

663 Although the violation function involves a max operation and is not twice differentiable at the kink  
 664 points of the max, we can approximate the max in practice using the softplus function, which is  
 665 infinitely differentiable and can approximate the max arbitrarily well. Therefore, this assumption  
 666 remains reasonable in practice due to the smooth approximation provided by the softplus function.  
 667 This assumption enables us to bound the bias caused by the truncation as in the following lemma.

**Lemma 2.** Suppose Assumption 4 holds and that the feasibility-seeking procedure is unrolled with the stepsize  $\eta_\phi \in (0, 1/L_\phi]$  for a total of  $K$  iterations, but gradient tracking is performed only for the first  $K' \in [K_0, K]$  iterations, with the remaining iterations treated as pass-through during backpropagation. Define the true Jacobian when all  $K$  iterations are tracked as  $J_y \text{FS}^K(y_\theta(x); x)$  and the truncated Jacobian as  $J_y^{\text{trun}} \text{FS}^K(y_\theta(x); x)$ :

$$J_y \text{FS}^K(y_\theta(x); x) = \frac{\partial s_K}{\partial s_0}, \quad J_y^{\text{trun}} \text{FS}^K(y_\theta(x); x) = \frac{\partial s_{K'}}{\partial s_0}$$

Then, the bias caused by the truncation satisfies

$$\left\| J_y \text{FS}^K(y_\theta(x); x) - J_y^{\text{trun}} \text{FS}^K(y_\theta(x); x) \right\|_2 \leq C(1 - \delta^{K-K'})\delta^{K'}$$

where  $\delta = 1 - \eta_\phi \mu' \in [0, 1)$  and  $C$  is a finite positive constant.

Since the truncation occurs during the backward pass, it also perturbs the stochastic gradient. We denote the resulting truncated gradient estimate by  $\tilde{\nabla}^{\text{trun}} L(\theta_t)$ . The next lemma quantifies this effect.

**Lemma 3.** For gradient estimate without truncation, suppose  $\mathbb{E}_t [\|\tilde{\nabla} L(\theta_t)\|_2^2] \leq G$  and  $\mathbb{E}_t [\tilde{\nabla} L(\theta_t)] = \nabla L(\theta_t)$  for all  $t$ . Then, when the gradient truncation is applied, under Assumptions in Lemma 2, the truncated gradient estimate satisfies

$$\begin{aligned} \mathbb{E}_t [\|\tilde{\nabla}^{\text{trun}} L(\theta_t)\|_2^2] &\leq G + \sigma_{K'}, \\ \mathbb{E}_t [\tilde{\nabla}^{\text{trun}} L(\theta_t)] &= \nabla L(\theta_t) + \varepsilon_{K'}. \end{aligned}$$

where  $\sigma_{K'} = \mathcal{O}(\delta^{K'})$  and  $\varepsilon_{K'} \in \mathbb{R}^{n_\theta}$  with  $\|\varepsilon_{K'}\| = \mathcal{O}(\delta^{K'})$ .

This lemma shows that the truncation introduces bounded biases to the variance and expectation of the gradient estimate, yet these biases decay exponentially fast in the truncation depth  $K'$ . This qualification allows us to understand the behavior of the SGD when the truncation is applied.

**Theorem 4.** Suppose Assumption 1, 2, 3, and the assumptions of Lemma 3 hold. Then, after running the SGD for  $T$  iterations with step size  $\eta = \eta_0/\sqrt{T}$  with  $\eta_0 > 0$ , there is at least one iteration  $t \in \{0, \dots, T-1\}$  satisfying

$$\mathbb{E} [\|\nabla L(\theta_t)\|_2^2] \leq \mathcal{O}(T^{-1/2} + \delta^{K'}), \quad (\text{D.9})$$

$$\mathbb{E} [\|\nabla \mathcal{L}(\theta_t)\|_2] \leq \mathcal{O}(T^{-1/4} + \gamma^K + \delta^{K'}), \quad (\text{D.10})$$

where  $\gamma$  and  $\delta$  are defined as in Theorem 4 and Lemma 2.

This theorem characterizes how the bias caused by truncating the gradient tracking in the feasibility-seeking procedure propagates to the convergence of SGD. We see that the upper bounds of the expected gradient norms contain bias terms of order  $\mathcal{O}(\gamma^K)$  and  $\mathcal{O}(\delta^{K'})$ . Since  $0 < \gamma, \delta \leq 1$ , even modest values of  $K$  and  $K'$  suffice to render these biases negligible in practice. We now present the proofs of the above lemmas and theorems.

## D.2.1 Proof of Lemma 2

Lemma 2 is an immediate result of the following lemma.

**Lemma 4.** Let  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$  be twice differentiable,  $L$ -smooth, satisfy the PL condition with constant  $\mu$ . Suppose there exists a neighborhood  $\mathcal{N}$  of  $s_K$  on which  $\varphi$  is strongly convex, i.e.,  $\nabla^2 \varphi(s) \succeq \mu' I, \forall s \in \mathcal{N}$  with  $\mu' > 0$ . Define

$$J^{\text{true}} = \frac{\partial s_K}{\partial s_0} = \prod_{k=0}^{K-1} (I - \eta \nabla^2 \varphi(s_k)), \quad J^{\text{trun}} = \frac{\partial s_{K'}}{\partial s_0} = \prod_{k=0}^{K'-1} (I - \eta \nabla^2 \varphi(s_k)).$$

If  $s_k \in \mathcal{N}$  for all  $k \geq K_0$  and stepsize of gradient descent is chosen  $\eta \in (0, 1/L]$ , then for any  $K' \in [K_0, K]$ , the error matrix  $E = J^{\text{true}} - J^{\text{trun}}$  satisfies

$$\|E\|_2 \leq C(1 - \delta^{K-K'})\delta^{K'},$$

where  $\delta = 1 - \eta \mu' \in [0, 1)$  and  $C = \frac{L}{\mu'} \delta^{-K_0} (1 + \eta L)^{K_0}$ .

702 *Proof.* Define  $J_k = \frac{\partial s_{k+1}}{\partial s_k} = I - \eta \nabla^2 \varphi(s_k)$  and rewrite the norm of error matrix  $E$  as

$$\begin{aligned} \|E\|_2 &= \|(J_{K-1} \dots J_{K'} - I)(J_{K'-1} \dots J_0)\|_2 \\ &\leq \|J_{K-1} \dots J_{K'} - I\|_2 \|J_{K'-1} \dots J_0\|_2. \end{aligned} \quad (\text{D.11})$$

703 **Step 1:**  $0 \leq k < K'$ .

704 For  $k < K_0$ , the  $L$ -smoothness gives  $-LI \preceq \nabla^2 \varphi(s_k) \preceq LI$ , so

$$\|J_k\|_2 = \|I - \eta \nabla^2 \varphi(s_k)\|_2 \leq 1 + \eta L.$$

705 For  $K_0 \leq k < K'$ , strong convexity implies  $\mu' I \preceq \nabla^2 \varphi(s_k) \preceq LI$ , and then we obtain

$$\|J_k\|_2 \leq 1 - \eta \mu'.$$

706 Let  $\delta := 1 - \eta \mu' \in [0, 1)$ . Then, we use the sub-multiplicativity property of the spectral norm to obtain

$$\|(J_{K'-1} \dots J_0)\|_2 \leq \|(J_{K'-1} \dots J_{K_0})\|_2 \|(J_{K_0-1} \dots J_0)\|_2 \leq \delta^{K'-K_0} (1 + \eta L)^{K_0}. \quad (\text{D.12})$$

708 **Step 2:**  $K' \leq k < K$ .

709 Use the telescoping sum for  $\|(J_{K-1} \dots J_{K'} - I)\|_2$ :

$$\begin{aligned} \|J_{K-1} \dots J_{K'} - I\|_2 &= \left\| \sum_{k=K'}^{K-1} \left( \prod_{j=k+1}^{K-1} J_j \right) (J_k - I) \right\|_2 \\ &\leq \sum_{k=K'}^{K-1} \left\| \prod_{j=k+1}^{K-1} J_j \right\|_2 \|J_k - I\|_2. \end{aligned}$$

710 From the  $L$ -smoothness we have  $\|J_k - I\|_2 = \|\eta \nabla^2 \varphi(s_k)\|_2 \leq \eta L$ . For  $k \geq K' > K_0$ , by  
711 assumption, we are in the neighborhood  $\mathcal{N}$  where  $\varphi$  is locally strongly convex, and thus we obtain

$$\left\| \prod_{j=k+1}^{K-1} J_j \right\|_2 \leq \delta^{K-k-1},$$

712 which results in

$$\|J_{K-1} \dots J_{K'} - I\|_2 \leq \eta L \sum_{k=K'}^{K-1} \delta^{K-k-1}.$$

713 We change the index of the summation to  $i = K - k - 1$ . When  $k = K'$ ,  $i = K - K' - 1$ . When  
714  $k = K - 1$ ,  $i = 0$ . The sum becomes a finite geometric series with  $0 \leq \delta < 1$ :

$$\begin{aligned} \|J_{K-1} \dots J_{K'} - I\|_2 &\leq \eta L \sum_{i=0}^{K-K'-1} \delta^i \\ &= \eta L \frac{1 - \delta^{K-K'}}{1 - \delta} \\ &= \frac{L}{\mu'} (1 - \delta^{K-K'}). \end{aligned} \quad (\text{D.13})$$

715 **Step 3:** Combining the bounds

716 Substituting (D.12) and (D.13) into the inequality of error matrix (D.11) yields

$$\|E\|_2 \leq \frac{L}{\mu'} (1 - \delta^{K-K'}) \delta^{K'-K_0} (1 + \eta L)^{K_0}.$$

717 Define the constant  $C = \frac{L}{\mu'} \delta^{-K_0} (1 + \eta L)^{K_0}$ . We can rewrite the bound as follows:

$$\|E\|_2 \leq C (1 - \delta^{K-K'}) \delta^{K'}.$$

718 □

719 **D.2.2 Proof of Lemma 3**

720 *Proof.* For a minibatch with  $D$  samples, we have

$$\begin{aligned}
& \left\| \tilde{\nabla} L(\theta_t) - \tilde{\nabla}^{\text{trun}} L(\theta_t) \right\|_2 \\
&= \left\| \frac{1}{D} \sum_{i=1}^D (J_{\theta} y_{\theta}(x^{(i)}))^{\top} \nabla_y F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}^K(x^{(i)})) \right. \\
&\quad + (J_{\theta} y_{\theta}(x^{(i)}))^{\top} (J_y \text{FS}^K(y_{\theta}(x^{(i)}); x^{(i)}))^{\top} \nabla_{\hat{y}} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}^K(x^{(i)})) \\
&\quad - (J_{\theta} y_{\theta}(x^{(i)}))^{\top} \nabla_y F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}^K(x^{(i)})) \\
&\quad \left. - (J_{\theta} y_{\theta}(x^{(i)}))^{\top} (J_y^{\text{trun}} \text{FS}^K(y_{\theta}(x^{(i)}); x^{(i)}))^{\top} \nabla_{\hat{y}} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}^K(x^{(i)})) \right\|_2 \\
&= \left\| \frac{1}{D} \sum_{i=1}^D (J_{\theta} y_{\theta}(x^{(i)}))^{\top} (J_y \text{FS}^K(y_{\theta}(x^{(i)}); x^{(i)})) \right. \\
&\quad \left. - J_y^{\text{trun}} \text{FS}^K(y_{\theta}(x^{(i)}); x^{(i)})^{\top} \nabla_{\hat{y}} F(y_{\theta}(x^{(i)}), \hat{y}_{\theta}^K(x^{(i)})) \right\|_2 \\
&\leq \frac{1}{D} \sum_{i=1}^D B_N B_F C (1 - \delta^{K-K'}) \delta^{K'} \quad (\text{by Lemma 2}) \\
&= B_N B_F C (1 - \delta^{K-K'}) \delta^{K'}. \tag{D.14}
\end{aligned}$$

721 Then, by triangle inequality, we have the following bound

$$\begin{aligned}
\mathbb{E}_t \left[ \left\| \tilde{\nabla}^{\text{trun}} L(\theta_t) \right\|_2^2 \right] &\leq \mathbb{E}_t \left[ \left( \left\| \tilde{\nabla}^{\text{trun}} L(\theta_t) - \tilde{\nabla} L(\theta_t) \right\|_2 + \left\| \tilde{\nabla} L(\theta_t) \right\|_2 \right)^2 \right] \\
&\leq B_N B_F C (1 - \delta^{K-K'})^2 \delta^{2K'} + G + 2B_N B_F C \sqrt{G} (1 - \delta^{K-K'}) \delta^{K'} \\
&= G + \mathcal{O}(\delta^{K'}).
\end{aligned}$$

722 In addition, by (D.14), we have that

$$\left\| \tilde{\nabla} L(\theta_t) - \tilde{\nabla}^{\text{trun}} L(\theta_t) \right\|_{\infty} \leq \left\| \tilde{\nabla} L(\theta_t) - \tilde{\nabla}^{\text{trun}} L(\theta_t) \right\|_2 \leq B_N B_F C (1 - \delta^{K-K'}) \delta^{K'} := v_{K'},$$

723 which results in

$$\tilde{\nabla} L(\theta_t) - v_{K'} \mathbf{1} \leq \tilde{\nabla}^{\text{trun}} L(\theta_t) \leq \tilde{\nabla} L(\theta_t) + v_{K'} \mathbf{1} \quad (\text{element-wise}).$$

724 There exists  $\varepsilon_{K'} \in \mathbb{R}^{n_{\theta}}$  such that  $\|\varepsilon_{K'}\|_2 \leq v_{K'} \sqrt{n_{\theta}}$  that satisfies

$$\mathbb{E}_t \left[ \tilde{\nabla}^{\text{trun}} L(\theta_t) \right] = \mathbb{E}_t \left[ \tilde{\nabla} L(\theta_t) \right] + \varepsilon_{K'} = \nabla L(\theta_t) + \varepsilon_{K'}.$$

725 Since  $v_{K'} = \mathcal{O}(\delta^{K'})$ , the proof is complete.  $\square$

726 **D.2.3 Proof of Theorem 4**

727 *Proof.* **Gradient bound for finite-unrolling loss with truncated gradient.** By the quadratic upper  
728 bound of the  $L$ -smooth function and bounded variance of the stochastic gradient (Lemma 3):

$$\begin{aligned}
\mathbb{E}_t [L(\theta_{t+1})] &\leq L(\theta_t) - \eta \langle \nabla L(\theta_t), \mathbb{E}_t [\tilde{\nabla}^{\text{trun}} L(\theta_t)] \rangle + \frac{L_N \eta^2}{2} \mathbb{E}_t \left[ \left\| \tilde{\nabla}^{\text{trun}} L(\theta_t) \right\|_2^2 \right] \\
&= L(\theta_t) - \eta \langle \nabla L(\theta_t), \nabla L(\theta_t) + \varepsilon_{K'} \rangle + \frac{L_N \eta^2}{2} (G + \sigma_{K'}). \tag{D.15}
\end{aligned}$$



729 We observe that

$$\begin{aligned}
\langle \nabla L(\theta_t), \nabla L(\theta_t) + \varepsilon_{K'} \rangle &= \|\nabla L(\theta_t)\|_2^2 + \langle \nabla L(\theta_t), \varepsilon_{K'} \rangle + \frac{1}{2} \|\varepsilon_{K'}\|_2^2 - \frac{1}{2} \|\varepsilon_{K'}\|_2^2 \\
&= \frac{1}{2} \|\nabla L(\theta_t)\|_2^2 - \frac{1}{2} \|\varepsilon_{K'}\|_2^2 + \frac{1}{2} (\|\nabla L(\theta_t)\|_2^2 + 2\langle \nabla L(\theta_t), \varepsilon_{K'} \rangle + \|\varepsilon_{K'}\|_2^2) \\
&= \frac{1}{2} \|\nabla L(\theta_t)\|_2^2 - \frac{1}{2} \|\varepsilon_{K'}\|_2^2 + \frac{1}{2} \|\nabla L(\theta_t) + \varepsilon_{K'}\|_2^2 \\
&\geq \frac{1}{2} \|\nabla L(\theta_t)\|_2^2 - \frac{1}{2} \|\varepsilon_{K'}\|_2^2.
\end{aligned}$$

730 Using this in (D.15) and rearranging terms, we obtain

$$\|\nabla L(\theta_t)\|_2^2 \leq \frac{2}{\eta} (L(\theta_t) - \mathbb{E}_t[L(\theta_{t+1})]) + \|\varepsilon_{K'}\|_2^2 + L_N \eta (G + \sigma_{K'}).$$

731 Taking the expectation for both sides and the average over  $T$  steps, we obtain

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla L(\theta_t)\|_2^2] &\leq \frac{2}{\eta T} \mathbb{E} \left[ \sum_{t=0}^{T-1} L(\theta_t) - L(\theta_{t+1}) \right] + \|\varepsilon_{K'}\|_2^2 + L_N \eta (G + \sigma_{K'}) \\
&\leq \frac{2}{\eta T} (L(\theta_0) - L^*) + \|\varepsilon_{K'}\|_2^2 + L_N \eta (G + \sigma_{K'}).
\end{aligned}$$

732 Since  $\min_{t \in \{0, \dots, T-1\}} \|\nabla L(\theta_t)\|_2^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla L(\theta_t)\|_2^2$ , we obtain the first statement of the  
733 theorem by replacing the stepsize  $\eta = \eta_0 / \sqrt{T}$ .

734 Using the same argument as the proof of Theorem 2, there exists positive constant  $\kappa_1, \kappa_2$  such that

$$\mathbb{E}[\|\nabla \mathcal{L}(\theta_t)\|_2] \leq \kappa_1 T^{-1/4} + \kappa_2 \gamma^{K-1} + \|\varepsilon_{K'}\|_2^2 + \frac{L_N \eta_0}{\sqrt{T}} \sigma_{K'}.$$

735

□