

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 TOWARDS A SHARP ANALYSIS OF LEARNING OF- FLINE $f$ -DIVERGENCE-REGULARIZED CONTEXTUAL BANDITS

Anonymous authors

Paper under double-blind review

## ABSTRACT

Many offline reinforcement learning algorithms are underpinned by  $f$ -divergence regularization, but their sample complexity *defined with respect to regularized objectives* still lacks tight analyses, especially in terms of concrete data coverage conditions. In this paper, we study the exact concentrability requirements to achieve the  $\tilde{\Theta}(\epsilon^{-1})$  sample complexity for offline  $f$ -divergence-regularized contextual bandits. For reverse Kullback–Leibler (KL) divergence, arguably the most commonly used one, we achieve an  $\tilde{O}(\epsilon^{-1})$  sample complexity under single-policy concentrability for the first time via a novel pessimism-based analysis, surpassing existing  $\tilde{O}(\epsilon^{-1})$  bound under all-policy concentrability and  $\tilde{O}(\epsilon^{-2})$  bound under single-policy concentrability. We also propose a near-matching lower bound, demonstrating that a multiplicative dependency on single-policy concentrability is necessary to maximally exploit the curvature property of reverse KL. Moreover, for  $f$ -divergences with strongly convex  $f$ , to which reverse KL *does not* belong, we show that the sharp sample complexity  $\tilde{\Theta}(\epsilon^{-1})$  is achievable even without pessimistic estimation or single-policy concentrability. We further corroborate our theoretical insights with numerical experiments and extend our analysis to contextual dueling bandits. We believe these results take a significant step towards a comprehensive understanding of objectives with  $f$ -divergence regularization.

## 1 INTRODUCTION

Due to the data-hungry and instable nature of reinforcement learning (RL), divergences that are straightforward to estimate via Monte Carlo or amenable to constrained optimization stand out from numerous candidates (Rényi, 1961; Csiszár, 1967; Müller, 1997; Basseville, 2013) as regularizers; the former family is typically  *$f$ -divergence* (Rényi, 1961) because any of them is an expectation, for which empirical average is a good proxy (Levine, 2018; Levine et al., 2020); and the latter class subsumes those with nice positive curvatures (e.g., Bregman divergence (Bregman, 1967) induced by strongly convex functions). In particular, *Kullback–Leibler (KL) divergence* is the only one at the intersection of  $f$ -divergence and Bregman divergence (Jiao et al., 2014, Theorem 5), indicating its theoretical advantage among common choices from both computational and statistical aspects. Also, the *KL-regularized RL objective* is arguably the most popular one in practice:

$$J(\pi) = \mathbb{E}_\pi[r] - \eta^{-1} \text{KL}(\pi \parallel \pi^{\text{ref}}), \quad (1.1)$$

where  $r$  is the reward,  $\pi^{\text{ref}}$  is a reference policy,  $\text{KL}(\pi \parallel \pi^{\text{ref}})$  is the reverse KL divergence, and  $\eta > 0$  is the inverse temperature. When  $\pi^{\text{ref}}$  is uniform, (1.1) reduces to the entropy-regularized objective that encourages diverse actions and enhances robustness (Williams, 1992; Ziebart et al., 2008; Levine & Koltun, 2013; Levine et al., 2016; Haarnoja et al., 2018; Richemond et al., 2024; Liu et al., 2024). KL regularization has also been widely used in the RL fine-tuning of large language models (Ouyang et al., 2022; Rafailov et al., 2023), where  $\pi^{\text{ref}}$  is the base model. Given its widespread use, there has been a surge of interest in understanding the role of KL regularization in RL by both empirical studies (Ahmed et al., 2019; Liu et al., 2019) and theoretical analysis (Geist et al., 2019; Vieillard et al., 2020; Kozuno et al., 2022). There are also lines of research on KL regularization in online learning (Cai et al., 2020; He et al., 2022; Ji et al., 2023) and convex optimization (Neu et al.,

054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1979  
1980  
1981  
1982  
1983  
1984  
1

108 the discrepancy between any two functions in the function class. To the best of our knowledge,  
 109 this machinery has not been used in the standard analysis of existing offline RL algorithms and  
 110 may be of independent interest.

111 • For  $f$ -divergence-regularized objectives with strongly convex  $f$ , we design a truly lightweight  
 112 algorithm free of pessimism-based gadgets and still obtain the  $\tilde{\Theta}(\epsilon^{-1})$  sample complexity certified  
 113 by a matching lower bound without coverage conditions.

114 • We verify the statistical rates above in numerical experiments, and demonstrate the versatility of  
 115 all algorithmic and constructive proof ideas above by extending them to  $f$ -divergence-regularized  
 116 contextual dueling bandits (CDBs), achieving similar  $\tilde{\Theta}(\epsilon^{-1})$  sample complexity bounds. More-  
 117 over, all algorithms are applicable for reward function classes with small metric entropy.

118

119 **1.2 KEY RELATED WORK**

120 We review two key lines of theoretical progress that are relevant to our algorithm design and analysis.

121 **Pessimism in offline RL.** The principle of pessimism has been underpinning offline RL for both  
 122 the tabular (Rashidinejad et al., 2021) and function approximation (Jin et al., 2021) settings under the  
 123 name of lower confidence bound (LCB). For contextual bandits, it is behind the adaptively optimal  
 124 sample complexity analysis (Li et al., 2022). Shi et al. (2022) proposed a LCB-based model-free  
 125 algorithm for tabular RL with near-optimal guarantee. Jin et al. (2021); Xiong et al. (2022); Di et al.  
 126 (2024) utilized LCB in conjunction with the classic least-square value iteration paradigm to derive  
 127  $\tilde{O}(\epsilon^{-2})$  sample complexity results for model-free RL with function approximation. The line of work  
 128 from Rashidinejad et al. (2021); Xie et al. (2021b) to Li et al. (2024) settled the sample complexity  
 129 of tabular model-based RL via pessimistic estimators exploiting the variance information. It is  
 130 also possible to leverage the idea of pessimism to design model-based algorithms under general  
 131 function approximation that are at least statistically efficient (Xie et al., 2021a; Uehara & Sun, 2021;  
 132 Wang et al., 2024). The principle of pessimism has also been applied in counterfactual empirical  
 133 risk minimization (Swaminathan & Joachims, 2015; London & Sandler, 2019) and offline policy  
 134 learning (Sakhi et al., 2023; 2024), which are orthogonal to our contributions.

135 However, in terms of risk decomposition, to the best of our knowledge, none of these pessimism-  
 136 based analyses really goes beyond the performance difference lemma (Foster & Rakhlin, 2023,  
 137 Lemma 13) or simulation lemma (Foster & Rakhlin, 2023, Lemma 23); both of which are not  
 138 able to capture the strong concavity of KL-regularized objectives even in the bandit setting. The  
 139 algorithmic idea of using pessimistic least-square estimators under general function approximation  
 140 in Jin et al. (2021); Di et al. (2024) is similar to ours, but their sub-optimality gap is bounded by the  
 141 sum of bonuses, which cannot directly lead to the desired sample complexity of our objective.

142 **Offline CDBs.** CDBs (Dudík et al., 2015) is the contextual extension of dueling bandits in the  
 143 classic literature of online learning from pairwise comparisons (Yue et al., 2012; Zoghi et al., 2014).  
 144 Since the empirical breakthrough of preference-based RL fine-tuning of LLMs (Ouyang et al., 2022),  
 145 the theory of offline CDBs has received more attention under linear function approximation (Zhu  
 146 et al., 2023; Xiong et al., 2024) and general function approximation (Zhan et al., 2022; Zhao et al.,  
 147 2024; Song et al., 2024; Huang et al., 2025b). Preference models without stochastic transitivity  
 148 (Munos et al., 2023; Ye et al., 2024; Wu et al., 2024; Zhang et al., 2024) are beyond the scope of  
 149 this work, namely, our preference labels are assumed to follow the Bradley-Terry Model (Bradley &  
 150 Terry, 1952).

151 **Notation.** The sets  $\mathcal{S}$  and  $\mathcal{A}$  are assumed to be countable throughout the paper. For nonnegative  
 152 sequences  $\{x_n\}$  and  $\{y_n\}$ , we write  $x_n = O(y_n)$  if  $\limsup_{n \rightarrow \infty} x_n/y_n < \infty$ ,  $y_n = \Omega(x_n)$  if  
 153  $x_n = O(y_n)$ , and  $y_n = \Theta(x_n)$  if  $x_n = O(y_n)$  and  $x_n = \Omega(y_n)$ . We further employ  $\tilde{O}(\cdot)$ ,  $\tilde{\Omega}(\cdot)$ ,  
 154 and  $\tilde{\Theta}$  to hide polylog factors. For countable  $\mathcal{X}$  and  $\mathcal{Y}$ , we denote the family of probability kernels  
 155 from  $\mathcal{X}$  to  $\mathcal{Y}$  by  $\Delta(\mathcal{Y}|\mathcal{X})$ . For  $g : \mathcal{X} \rightarrow \mathbb{R}$ , its infinity norm is denoted by  $\|g\|_\infty := \sup_{x \in \mathcal{X}} |g(x)|$ .  
 156 For a pair of probability measures  $P \ll Q$  on the same space and function  $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ , their  
 157  $f$ -divergence is  $D_f(P||Q) := \int f(dP/dQ) dQ$ . Specifically, when  $f(x) = x \log x$ ,  $f$ -divergence  
 158 becomes KL divergence denoted as  $\text{KL}(P||Q) := \int \log(dP/dQ) dP$ , and when  $f(x) = |x - 1|/2$ ,  
 159 it becomes the total variation (TV) distance, which is denoted as  $\text{TV}(P||Q) := 0.5 \int |dP - dQ|$ .  
 160 We use  $\text{supp}(P)$  to denote the support set of  $P$ .

## 162 2 KL-REGULARIZED CONTEXTUAL BANDITS

164 In this section, we introduce a pessimism-based algorithm, PCB-KL, for offline KL-regularized  
 165 contextual bandits. We then showcase our novel analysis techniques for PCB-KL, which couples  
 166 the algorithmic pessimism with the curvature property of KL-regularized objectives.

### 167 2.1 PROBLEM SETUP

169 We consider contextual bandit, which is denoted by a tuple  $(\mathcal{S}, \mathcal{A}, r, \pi^{\text{ref}})$ . Specifically,  $\mathcal{S}$  is the  
 170 context space,  $\mathcal{A}$  is the action space and  $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$  is the reward function. In the offline  
 171 setting, the agent only has access to an i.i.d. dataset  $\mathcal{D} = \{(s_i, a_i, r_i)\}_{i=1}^n$ . Here  $s_i$ s are states  
 172 sampled from  $\rho \in \Delta(\mathcal{S})$ ,  $a_i \in \mathcal{A}$  is the action taken from a *behavior policy*, and  $r_i$  is the observed  
 173 reward given by  $r_i = r(s_i, a_i) + \varepsilon_i$ , where  $\varepsilon_i$  is 1-sub-Gaussian (Lattimore & Szepesvári, 2020,  
 174 Definition 5.2). In this work, we consider the *KL-regularized objective*

$$175 \quad J_\eta(\pi) := \mathbb{E}_{(s, a) \sim \rho \times \pi} \left[ r(s, a) - \eta^{-1} \log \frac{\pi(a|s)}{\pi^{\text{ref}}(a|s)} \right], \quad (2.1)$$

178 where  $\pi^{\text{ref}}$  is a known reference policy and the “inverse temperature”  $\eta$  controls the intensity of reg-  
 179 ularization. For simplicity, we assume that  $\pi^{\text{ref}}$  is also the behavior policy that generates the dataset  
 180  $\mathcal{D}$ , which is similar to the type of “behavior regularization” studied in Zhan et al. (2022). The unique  
 181 optimal policy  $\pi_\eta^* := \text{argmax}_{\pi \in \Delta(\mathcal{A}|\mathcal{S})} J_\eta(\pi)$  is given by (See, e.g., Zhang 2023, Proposition 7.16)<sup>3</sup>

$$182 \quad \pi^*(\cdot|s) \propto \pi^{\text{ref}}(\cdot|s) \exp(\eta \cdot r(s, \cdot)), \forall s \in \mathcal{S}. \quad (2.2)$$

184 A policy  $\pi$  is said to be  $\epsilon$ -optimal if  $\text{SubOpt}_{\text{RKL}}(\pi) := J(\pi^*) - J(\pi) \leq \epsilon$  and the goal of the  
 185 agent is to find one such policy using  $\mathcal{D}$ . Note that  $\text{SubOpt}_{\text{RKL}}(\cdot)$  is defined through (2.1) and thus  
 186 **depends on**  $\eta$ . To ensure that  $\epsilon$ -optimality is achievable, we assume that  $r$  lies in a known function  
 187 class  $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \rightarrow [0, 1])$ , from which the agent obtains an estimator  $\hat{r}$ . More specifically, we  
 188 work with general function approximation under realizability, which is as follows.

189 **Assumption 2.1.** For this known function class  $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \rightarrow [0, 1])$ ,  $\exists g^* \in \mathcal{G}$  with  $g^* = r$ .

191 We also employ the standard notion of covering number (Wainwright, 2019, Definition 5.1) as the  
 192 complexity measure of the reward function class  $\mathcal{G}$ .

193 **Definition 2.2** ( $\epsilon$ -net and covering number). Given a function class  $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R})$ , a finite  
 194 set  $\mathcal{G}(\epsilon) \subset \mathcal{G}$  is an  $\epsilon$ -net of  $\mathcal{G}$  w.r.t.  $\|\cdot\|_\infty$ , if for any  $g \in \mathcal{G}$ , there exists  $g' \in \mathcal{G}(\epsilon)$  such that  
 195  $\|g - g'\|_\infty \leq \epsilon$ . The  $\epsilon$ -covering number is the smallest cardinality  $\mathcal{N}_{\mathcal{G}}(\epsilon)$  of such  $\mathcal{G}(\epsilon)$ .

196 **Assumption 2.3.** For any  $\epsilon_c > 0$ , the  $\epsilon_c$ -covering number  $\mathcal{N}_{\mathcal{G}}(\epsilon_c)$  of  $\mathcal{G}$  is  $\text{poly}(\epsilon_c^{-1})$ .

198 Assumption 2.3 allowing  $\log \mathcal{N}_{\mathcal{G}}(\epsilon)$  to be roughly negligible is arguably mild. For example, when  
 199  $\mathcal{G}$  is the class of linear functions of dimension  $d$  and radius  $R$ , the covering number is  $\mathcal{N}_{\mathcal{G}}(\epsilon) =$   
 200  $O((1 + R\epsilon^{-1})^d)$  (Jin et al., 2020, Lemma D.6), which satisfies Assumption 2.3.

201 **Concentrability.** The data quality of  $\mathcal{D}$  collected by  $\pi^{\text{ref}}$  is typically characterized by *concentrability* in offline RL (Farahmand et al., 2010; Chen & Jiang, 2019; Jiang & Xie, 2024), which  
 202 quantifies the ability of the behavioral policy to generate diverse actions. We first define the density-  
 203 ratio-based concentrability as follows.

206 **Definition 2.4** (*Density-ratio-based* concentrability). For policy class  $\Pi$ , reference policy  $\pi^{\text{ref}}$ , the density-ratio-based all-policy concentrability  $C^\Pi$  is  $C^\Pi := \sup_{\pi \in \Pi, s \in \mathcal{S}, a \in \mathcal{A}} \pi(a|s)/\pi^{\text{ref}}(a|s)$ , whose single-policy counterpart under the optimal policy  
 207  $\pi^*$  is  $C^{\pi^*} := \sup_{s \in \mathcal{S}, a \in \mathcal{A}} \pi^*(a|s)/\pi^{\text{ref}}(a|s)$ .

210 In the definition above, small all-policy concentrability intuitively corresponds to  $\text{supp}(\pi^{\text{ref}})$  covering  
 211 all possible inputs. On the other hand, small single-policy concentrability means that  $\text{supp}(\pi^{\text{ref}})$   
 212 only subsumes  $\text{supp}(\pi^*)$ . In this paper, in addition to density-ratio-based concentrability, we also  
 213 adopt the following  $D^2$ -based concentrabilites to better capturing the nature of function class  $\mathcal{G}$ . In  
 214 detail, we start with the  $D^2$ -divergence as follows.

215 <sup>3</sup>We suppress  $J_\eta$  into  $J$  and  $\pi_\eta^*$  into  $\pi^*$  when they are clear in context in the following presentation.

216 **Definition 2.5.** Given a function class  $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R})$  and a fixed policy  $\pi$ , define the  $D^2$ -  
 217 divergence  $D_{\mathcal{G}}^2((s, a); \pi)$  as  
 218

$$219 \quad \sup_{g, h \in \mathcal{G}} \frac{(g(s, a) - h(s, a))^2}{220 \mathbb{E}_{(s', a') \sim \rho \times \pi} [(g(s', a') - h(s', a'))^2]}.$$

221 The ‘‘eluder dimension’’-type Definition 2.5 is directly inspired by Di et al. (2024); Zhao et al.  
 222 (2024), the intuition behind which is that given  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , a small  $D^2$ -divergence indicates  
 223 that for two functions  $g$  and  $h$ , if they are close under the behavior policy  $\pi$ , then they will also  
 224 be close on such pair  $(s, a)$ . Therefore, the  $D^2$ -divergence quantifies how well the estimation on  
 225 dataset collected by the behavior policy  $\pi$  can be generalized to a specific state-action pair.  
 226

227 **Remark 2.6.** For the tabular setting, a direct computation yields  $D^2(s, a) = (\rho(s)\pi^{\text{ref}}(a|s))^{-1}$ ,  
 228 which can be estimated by the visitation frequency empirically. Under linear function approxima-  
 229 tion, it is well known that  $D^2(s, a) = \|\phi(s, a)\|_{\Sigma^{-1}}^2$  under mild conditions of the parameter space,  
 230 where  $\Sigma = \mathbb{E}_{\rho \times \pi^{\text{ref}}} \phi(s, a)\phi(s, a)^\top$  is the covariance matrix, which can be estimated by empirical  
 231 covariance matrices in practice, potentially with ridge regularization. For more general function  
 232 classes like neural networks, the  $D^2$  can also be efficiently approximated by heuristics as discussed  
 233 in Xiong et al. (2024); Gupta et al. (2024); Xu et al. (2025).

234 We are now ready to define the two notions of concentrability conditions.

235 **Assumption 2.7** (All-policy concentrability). Given a reference policy  $\pi^{\text{ref}}$ , there exists  $D < \infty$   
 236 such that  $D^2 = \sup_{(s, a) \in \mathcal{S} \times \mathcal{A}} D_{\mathcal{G}}^2((s, a); \pi^{\text{ref}})$ .

237 Assumption 2.7 indicates that the errors on any state-action pairs can be bounded by the error on the  
 238 samples from  $\rho \times \pi$  up to a factor  $D$ , whose relaxed counterpart under the same  $\pi^{\text{ref}}$  is as follows.

239 **Assumption 2.8** (Single-policy concentrability).  $D_{\pi^*}^2 := \mathbb{E}_{(s, a) \sim \rho \times \pi^*} D_{\mathcal{G}}^2((s, a); \pi^{\text{ref}}) < \infty$ .

240 Assumption 2.8 indicates that the errors on the distributions of state-action pairs  $\rho \times \pi^*$  can be  
 241 bounded by the error on the samples from  $\rho \times \pi^{\text{ref}}$  up to some constant. For both types, the single-  
 242 policy concentrability assumption is strictly weaker than the all-policy concentrability assumption.  
 243 However, in general, the two quantities characterizing single-policy concentrability  $C^{\pi^*}$  and  $D_{\pi^*}^2$ ,  
 244 cannot be bounded by each other up to constant factors. In particular, we have  $D_{\pi^*}^2 \leq |\mathcal{S}||\mathcal{A}|C^{\pi^*}$ ,  
 245 indicating that  $C^{\pi^*}$  subsumes  $D_{\pi^*}^2$  when  $|\mathcal{S}|$  and  $|\mathcal{A}|$  can be seen as constants. [We refer the reader to Appendix B for a further discussion on the relation between  \$C^{\pi^\*}\$  and  \$D\_{\pi^\*}^2\$ .](#)

## 2.2 ALGORITHM

250 In this subsection, we present an offline bandit algorithm, KL-PCB, for KL-regularized contextual  
 251 bandits in Algorithm 1. KL-PCB first leverages least-square estimator to find a function  $\bar{g} \in \mathcal{G}$   
 252 that minimizes its risk on the offline dataset. In Zhao et al. (2024), such  $\bar{g}$  is directly applied to  
 253 construct the estimated policy. In contrast, we construct a pessimistic estimator of  $g^*$  following the  
 254 well-known pessimism principle in offline RL (Jin et al., 2021). Specifically, we define the bonus  
 255 term  $\Gamma_n$  through the confidence radius  $\beta = \sqrt{128 \log(2\mathcal{N}_{\mathcal{G}}(\epsilon)/\delta)/3n + 18\epsilon}$  as

$$256 \quad \Gamma_n(s, a) = \beta D_{\mathcal{G}}((s, a), \pi^{\text{ref}}), \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \quad (2.3)$$

257 We then obtain our pessimistic estimation  $\hat{g}$  by setting  $\hat{g} = \bar{g} - \Gamma_n$ , which is less than  $g^*$  with high  
 258 probability. Formally, let the event  $\mathcal{E}(\delta)$  given  $\delta > 0$  defined as

$$259 \quad \mathcal{E}(\delta) := \left\{ \sup_{(s, a) \in \mathcal{S} \times \mathcal{A}} \left[ |\bar{g} - g^*| - \Gamma_n \right] (s, a) \leq 0 \right\}, \quad (2.4)$$

260 on which the least square estimation  $\bar{g}$  obtained in Line 1 of Algorithm 1 does not deviate too much  
 261 from the true function  $g^*$  and therefore  $\hat{g}$  is a pessimistic estimation of  $g^*$ . We have the following  
 262 lemma indicating that this event holds with high probability.

263 **Lemma 2.9.** For all  $\delta > 0$ ,  $\mathcal{E}(\delta)$  holds with probability at least  $1 - \delta$ .

264 After obtaining the pessimistic estimation, KL-PCB output the policy  $\hat{\pi}$ , which maximizes the esti-  
 265 mated objective

$$266 \quad \hat{J}(\pi) = \mathbb{E}_{(s, a) \sim \rho \times \pi} \left[ \hat{g}(s, a) - \eta^{-1} \log \frac{\pi(a|s)}{\pi^{\text{ref}}(a|s)} \right],$$

---

**270 Algorithm 1 Offline KL-Regularized Pessimistic Contextual Bandits (KL-PCB)**


---

**271 Require:** regularization  $\eta$ , reference policy  $\pi^{\text{ref}}$ , offline dataset  $\mathcal{D}$ , function class  $\mathcal{G}$ 
**272 1:** Least square estimation of reward function  $\bar{g} \in \operatorname{argmin}_{g \in \mathcal{G}} \sum_{(s_i, a_i, r_i) \in \mathcal{D}} (g(s_i, a_i) - r_i)^2$ 
**273 2:** Let  $\hat{g} \leftarrow \bar{g} - \Gamma_n$ , where  $\Gamma_n$  is the bonus term in (2.3)

**274 Ensure:**  $\hat{\pi}(a|s) \propto \pi^{\text{ref}}(a|s) \exp(\eta \cdot \hat{g}(s, a))$ 


---

**278** the maximizer of which is the counterpart of (2.2), i.e.,

$$279 \hat{\pi}(a|s) \propto \pi^{\text{ref}}(a|s) \exp(\eta \cdot \hat{g}(s, a)).$$

**281 2.3 THEORETICAL RESULTS**
**283** The sample complexity for KL-regularized contextual bandits is settled in this subsection. We first  
**284** give the upper bound of KL-PCB.

**285 Theorem 2.10.** Under Assumption 2.8, for sufficiently small  $\epsilon \in (0, 1)$ , if we set  $\Gamma_n$  as in (2.3),  
**286** then  $n = \tilde{O}(\eta D_{\pi^*}^2 \epsilon^{-1} \log \mathcal{N}_{\mathcal{G}}(\epsilon))$  suffices to guarantee the output policy  $\hat{\pi}$  of Algorithm 1 to be  
**287**  $\epsilon$ -optimal with probability at least  $1 - \delta$ .

**289** Previously, Zhao et al. (2024) achieved an  $\tilde{O}(\epsilon^{-1})$  sample complexity under Assumption 2.7. As  
**290** a comparison, KL-PCB achieves the same  $\tilde{O}(\epsilon^{-1})$  sample complexity but only requiring Assump-  
**291** tion 2.8, which is weaker than Assumption 2.7. We also provide the sample complexity lower bound  
**292** of KL-regularized contextual bandits in the following theorem, which, together with Theorem 2.10,  
**293** demonstrates that single-policy concentrability is both necessary and sufficient for near-optimal of-  
**294** fline learning evaluated by KL-regularized objectives.

**295 Theorem 2.11.** For  $\forall S \geq 1$ ,  $\eta > 4 \log 2$ ,  $C^* \in (2, \exp(\eta/4)]$ , and any algorithm  
**296**  $\text{Alg}$ , there is a KL-regularized contextual bandit with  $C^{\pi^*} \leq C^*$  such that  $\text{Alg}$  requires  
**297**  $\Omega(\min\{\eta\epsilon^{-1}, \epsilon^{-2}\}C^* \log \mathcal{N}_{\mathcal{G}}(\epsilon))$  samples to find an  $\epsilon$ -optimal policy for sufficiently small  $\epsilon$ .

**299** Previously, Zhao et al. (2024) provided a sample complexity lower bound of  $\Omega(\eta \log \mathcal{N}_{\mathcal{G}}(\epsilon)/\epsilon)$  under  
**300** KL regularization. Foster et al. (2025) also provided a lower bound of  $\Omega(C^{\pi^*})$  for KL-regularized  
**301** objective to show the necessity of coverage. Compared to their results, our result shows that the  
**302** *multiplicative* dependency on  $C^{\pi^*}$  is necessary for the first time.

**303 Remark 2.12.** Theorem 2.11 shows that when  $\epsilon$  is sufficiently small, any algorithm for offline  
**304** KL-regularized contextual bandits requires at least  $\Omega(\eta C^{\pi^*})\epsilon^{-1} \log \mathcal{N}_{\mathcal{G}}(\epsilon)$  samples to output an  
**305**  $\epsilon$ -optimal policy. The presence of  $\exp(\text{poly}(\eta))$  in the range of  $C^*$  is inevitable, since we always  
**306** have  $C^{\pi^*} \leq \exp(\eta)$  in reverse KL regularized bandits with bounded rewards.

**307 Remark 2.13.** As discussed before, we might have some easy instances with  $D_{\pi^*}^2 \leq C^{\pi^*}$ ,  
**308** where KL-PCB outperforms the lower bound. This does not violates Theorem 2.11 since The-  
**309** orem 2.11 only guarantees that *there exist* some hard instances that all algorithms require at least  
**310**  $\Omega(\min\{\eta\epsilon^{-1}, \epsilon^{-2}\}C^* \log \mathcal{N}_{\mathcal{G}}(\epsilon))$  samples.

**312 2.4 PROOF OVERVIEW OF THEOREM 2.10**
**313** In this section, we summarize the novel techniques in the proof of Theorem 2.10, which is de-  
**314** ferred to Appendix D.4. At a high level, if we consider the regularized objective (1.1) multi-arm  
**315** bandits, then  $P \mapsto \text{KL}(P\|Q)$  is 1-strongly convex w.r.t.  $\text{TV}(\cdot\|\cdot)$  (Polyanskiy & Wu, 2025, Exer-  
**316** cise I.37), and thus  $J(\pi)$  is strongly concave. Therefore,  $J(\pi^*) - J(\hat{\pi})$  is possible to be of the order  
**317**  $[\text{TV}(\pi^*\|\hat{\pi})]^2 \approx \tilde{O}(n^{-1})$ , pretending that  $\pi^*$  is the unconstrained maximizer. In detail, we follow  
**318** the regret decomposition in Zhao et al. (2024), which is encompassed by the following lemma.

**319 Lemma 2.14.** Let  $g : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  be any reward function, then there exist some  $\gamma \in [0, 1]$  such  
**320** that the sub-optimality gap of  $\pi_g(\cdot|s) \propto \pi^{\text{ref}}(\cdot|s) \exp(\eta g(s, \cdot))$  can be bounded as

$$322 J(\pi^*) - J(\pi_g) \leq \eta \mathbb{E}_{(s, a) \sim \rho \times \pi_\gamma} [(g^* - g)^2(s, a)],$$

**323** where  $g_\gamma := \gamma g + (1 - \gamma)g^*$  and  $\pi_\gamma(\cdot|s) \propto \pi^{\text{ref}}(\cdot|s) \exp(\eta g_\gamma(s, \cdot))$ .

In Zhao et al. (2024), because the  $g$  in Lemma 2.14 is substituted with only the least-square estimator  $\bar{g}$  with no extra structures, the reliance on the “mid-point” policy  $\pi_\gamma$  can only be controlled all-policy concentrability. However, our  $g$  is the pessimistic estimator  $\hat{g}$  of  $g^*$  in Algorithm 1, and thus the presence of  $\pi_\gamma$  can be eliminated for free: let  $G(\gamma) := \mathbb{E}_{\rho \times \pi_\gamma} \left[ (\hat{g} - g^*)^2(s, a) \right]$  and  $\Delta(s, a) := (\hat{g} - g^*)(s, a) \leq 0$ , then a direct computation (detailed in the proof of Lemma D.3) yields

$$G'(\gamma) = \eta \mathbb{E}_\rho \left[ \mathbb{E}_{\pi_\gamma} [\Delta^3(s, a)] - \mathbb{E}_{\pi_\gamma} [\Delta^2(s, a)] \mathbb{E}_{\pi_\gamma} [\Delta(s, a)] \right] \leq 0. \quad (2.5)$$

This gives  $J(\pi^*) - J(\hat{\pi}) \leq \eta \mathbb{E}_{\rho \times \pi^*} [(\hat{g} - g^*)^2(s, a)]$ , which can be bounded with single-policy concentrability while still achieves the sharp dependency  $\epsilon^{-1}$  on  $\epsilon$ . Here, (2.5) holds due to a moment-based machinery in Lemma 2.15.

**Lemma 2.15.** If  $\mathbb{P}(X \leq 0) = 1$  and  $\mathbb{E}|X|^3 < \infty$ , then  $\mathbb{E}[X^3] - \mathbb{E}[X^2]\mathbb{E}[X] \leq 0$ .

The intuition behind Lemma 2.15 is natural:  $X$  and  $X^2$  cannot be positively correlated. Moreover, to the best of our knowledge, we are the first to unveil this moment-based structure in our non-standard pessimism-based analysis, from which the sharp upper bound follows. While pessimism is widely adopted to derive near-optimal statistical rates under single-policy concentrability in offline RL with reward maximization as the goal (See, e.g., Jin et al. (2021); Xiong et al. (2022)), the standard pessimism-based pipeline is not sharp enough for bounding the  $\text{SubOpt}_{\text{RKL}}(\hat{\pi})$  defined through *regularized objectives*, the reason of which is detailed in the last paragraph of Appendix A.1.

### 3 $f$ -DIVERGENCE-REGULARIZED CONTEXTUAL BANDITS

As discussed in Section 2, the fast rate implied by Theorems 2.10 and 2.11 is primarily achieved due to the strong convexity of  $\pi \mapsto \text{KL}(\pi \parallel \pi^{\text{ref}})$ . However, KL is just an instance of  $f$ -divergence with  $f(x) = x \log x$ , which is only locally strongly convex but not strongly convex. Motivated by this observation, we further examine  $f$ -divergence regularization with strongly convex  $f$ , which may introduce a more favorable curvature in the performance metric of offline learning in principle.

#### 3.1 PROBLEM SETUP

We study a contextual bandit setting similar to that in Section 2.1. In this section, we consider the following  $f$ -divergence regularized objective

$$J_{\eta, D_f}(\pi) := \mathbb{E}_{(s, a) \sim \rho \times \pi} [r(s, a)] - \eta^{-1} \mathbb{E}_{s \sim \rho} [D_f(\pi(\cdot | s) \parallel \pi^{\text{ref}}(\cdot | s))], \quad (3.1)$$

where  $\eta$  is the regularization intensity and  $D_f(p \parallel q) := \mathbb{E}_{a \sim q} [f(p(a)/q(a))]$  is the  $f$ -divergence. Let the optimal policy be  $\pi_{\eta, D_f}^* := \text{argmax}_{\pi \in \Delta(\mathcal{A} | \mathcal{S})} J_{\eta, D_f}(\pi)$  and we re-define the learning objective as searching for a policy  $\pi$  with  $\text{SubOpt}_{f \text{div}}(\pi) := J(\pi^*) - J(\pi) \leq \epsilon$ .<sup>4</sup> We consider those functions  $f : (0, +\infty) \rightarrow \mathbb{R}$  with a nice positive curvature condition in Assumption 3.1.

**Assumption 3.1.**  $f$  is  $\alpha$ -strongly convex, twice continuously differentiable, and  $f(1) = 0$ .

Many elementary functions like quadratic polynomials naturally satisfy Assumption 3.1. For instance, the 1-strongly convex  $f(x) = (x - 1)^2/2$  induces  $D_f(P \parallel Q) = \chi^2(P \parallel Q)$ , which is the  $\chi^2$ -divergence recently considered in RL literature (see e.g., Zhan et al. (2022); Huang et al. (2025b); Amortila et al. (2024)). This regularization exhibits a promising theoretical potential for relaxing the data coverage requirement for efficient offline policy learning (Huang et al., 2025b) and to be effective in preventing reward hacking (Laidlaw et al., 2025) against unregularized objectives. These favorable benefits are primary due to the observation that strongly convex  $f$ ’s impose a stronger penalization on actions out of the coverage of  $\pi^{\text{ref}}$ .

#### 3.2 ALGORITHM AND MAIN RESULTS

In this subsection, we present an offline learning algorithm for  $f$ -divergence regularized bandit,  $f$ -CB, in Algorithm 2. Algorithm 2 first leverages least-square estimator to find a function  $\bar{g} \in \mathcal{G}$  that minimizes its risk on the offline dataset. The algorithm then uses the least squares estimation  $\bar{g}$  to construct the output policy  $\hat{\pi}$ . Compared to Algorithm 1,  $f$ -CB does not require any procedure to construct pessimistic reward estimation, whose sample complexity upper bound is given as follows.

<sup>4</sup>We again suppress  $J_{\eta, D_f}(\cdot)$  into  $J(\cdot)$  and  $\pi_{\eta, D_f}^*$  into  $\pi^*$  when there is no confusion.

---

**Algorithm 2** Offline  $f$ -divergence Regularized Contextual Bandits ( $f$ -CB)

---

**Require:** regularization  $\eta$ , reference policy  $\pi^{\text{ref}}$ , function class  $\mathcal{G}$ , offline dataset  $\mathcal{D}$ 
1: Least square estimation  $\bar{g} \in \operatorname{argmin}_{g \in \mathcal{G}} \sum_{(s_i, a_i, r_i) \in \mathcal{D}} (g(s_i, a_i) - r_i)^2$ 2: Compute the optimal policy under the least-square reward estimator  $\bar{g}$  for  $s \in \mathcal{S}$  as

$$\hat{\pi}(\cdot|s) \leftarrow \operatorname{argmax}_{\pi(\cdot|s) \in \Delta(\mathcal{A})} \langle \pi(\cdot|s), \bar{g}(s, \cdot) \rangle + \eta^{-1} D_f(\pi(\cdot|s) \| \pi^{\text{ref}}(\cdot|s))$$

**Ensure:**  $\hat{\pi}$ 


---

**Theorem 3.2.** Under Assumption 3.1, for sufficiently small  $\epsilon \in (0, 1)$ , with probability at least  $1 - \delta$ ,  $n = \tilde{O}(\alpha^{-1} \eta \epsilon^{-1} \log \mathcal{N}_{\mathcal{G}}(\epsilon))$  is sufficient to guarantee the output policy  $\hat{\pi}$  of  $f$ -CB to be  $\epsilon$ -optimal.

**Remark 3.3.** Compared to the  $D_{\pi^*}^2$  dependency in Theorem 2.10, Theorem 3.2 shows that the sample complexity of Algorithm 2 gets rid of the dependency on any data coverage conditions when  $f$  is strongly convex. Intuitively, this is because the  $f$ -divergence regularization in this case is much stronger, so that both  $\pi^*$  and  $\hat{\pi}$  are close enough to  $\pi^{\text{ref}}$ .

The following hardness result justify the near-optimality of Theorem 3.2 for  $f$ -divergence-regularized contextual bandits.

**Theorem 3.4.** For any  $\epsilon \in (0, 1)$ ,  $\alpha > 0$ ,  $\eta > 0$ ,  $S > 32/3 \cdot \log 2$ , sufficiently small  $\epsilon$ , and algorithm  $\text{Alg}$ , there is an  $\alpha$ -strongly-convex function  $f$  and an  $f$ -divergence-regularized contextual bandit instance such that  $\text{Alg}$  requires at least  $\Omega(\alpha^{-1} \eta \epsilon^{-1} \log \mathcal{N}_{\mathcal{G}}(\epsilon))$  samples to return an  $\epsilon$ -optimal policy.

### 401 3.3 PROOF OVERVIEW OF THEOREM 3.2

We provide an overview of key analysis techniques for proving Theorem 3.2. Unlike KL-regularization, the  $\pi^*$  under  $f$ -divergence might not have a closed form. This means that the proof of Lemma 2.14, which relies on the closed form of  $\pi^*$ , cannot be directly adopted. Therefore, we address this from a dual-Bregman perspective. For the simplicity of presentation, we consider multi-armed bandits here and omit the subscript for context  $s$ .

We consider the function  $H(\pi) = \eta^{-1} D_f(\pi \| \pi^{\text{ref}})$ , which is the regularizer in the objective. Then its convex conjugate is given by  $H^*(r) = \sup_{\pi \in \Delta^d} \{ \langle \pi, r \rangle - H(\pi) \}$ , which is exactly the expected reward obtained by the optimal policy given reward function  $r$ . One observation is that when  $f$  is strongly convex, the induced  $f$ -divergence, and therefore the function  $H$  are also strongly convex. Therefore, let  $\pi_r = \operatorname{argmax}_{\pi} \{ \langle \pi, r \rangle - H_s(\pi) \}$  given some reward function  $r$ , the strong convexity of  $H(\pi)$  gives that  $\nabla H^*(r) = \pi_r$ . This leads to the following regret decomposition, which is one of our key observations:

$$\begin{aligned} J(\pi^*) - J(\hat{\pi}) &= \mathbb{E}_{a \sim \pi^*} [g^*(a)] - \mathbb{E}_{a \sim \hat{\pi}} [g^*(a)] - \eta^{-1} [D_f(\pi^* \| \pi^{\text{ref}}) - D_f(\hat{\pi} \| \pi^{\text{ref}})] \\ &= H^*(g^*) - H^*(\bar{g}) - \langle \hat{\pi}, g^* - \bar{g} \rangle \\ &= H^*(g^*) - H^*(\bar{g}) - \langle \nabla H^*(\bar{g}), g^* - \bar{g} \rangle, \end{aligned}$$

which is the Bregman divergence of the dual function  $H^*$  and therefore can be bounded by  $(g^* - \bar{g})^\top \nabla^2 H^*(\bar{g})(g^* - \bar{g})$  for some  $\bar{g}$ . By Proposition 3.2 in Penot (1994), when  $H$  is strongly convex, we can bound  $\nabla^2 H^*(\bar{g})$  as follows

$$\nabla^2 H^*(\bar{g}) \preceq (\nabla^2 H(\pi_{\bar{g}}))^{-1} \preceq \alpha^{-1} \eta \operatorname{diag}(\pi^{\text{ref}}(a_1), \dots, \pi^{\text{ref}}(a_{|\mathcal{A}|})),$$

which enables us to bound  $(g^* - \bar{g})^\top \nabla^2 H^*(\bar{g})(g^* - \bar{g})$  by  $\alpha^{-1} \eta \mathbb{E}_{\pi^{\text{ref}}} [(g^* - \bar{g})^2]$ . Since  $\mathbb{E}_{\pi^{\text{ref}}} [(g^* - \bar{g})^2]$  is not related to  $\pi^*$ , the upper bound is independent of any notion of concentrability.

## 4 EXPERIMENTS

### 4.1 TABULAR SETTING

We empirically check in this section the correctness of our matching bounds for KL and  $f$ -divergence on the simplest testbed: *two-armed* bandits, i.e.,  $\mathcal{A} = \{0, 1\}$ . We use one hard instance constructed in the proof of Theorem 2.11 (Appendix D.5) for the simulation under KL and

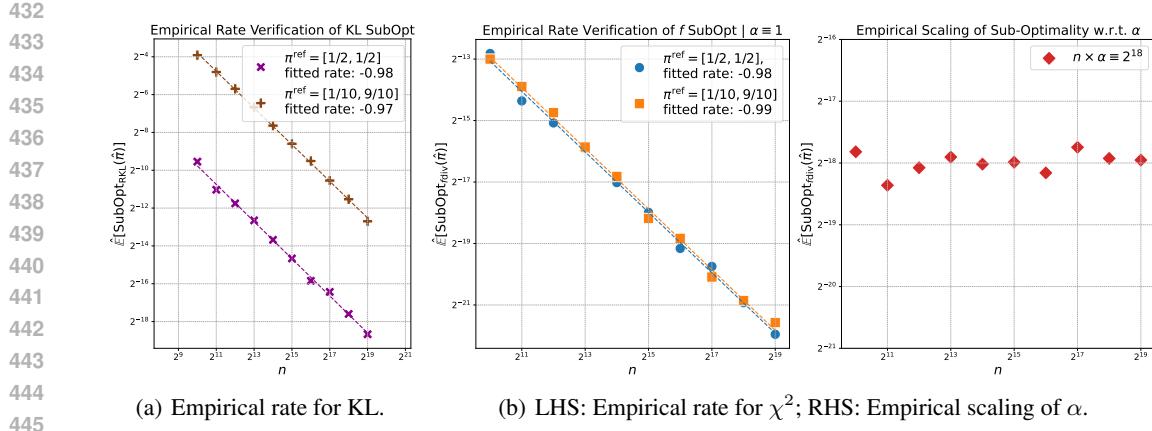


Figure 1: The empirical relation between  $\log_2 n$  and  $\log_2 \text{SubOpt}$ . The *fitted rate* means the slope of  $\log_2 n \sim \log_2 \text{SubOpt}$  estimated via linear regression. Here  $n$  is the sample size. Every point is the average over **100** independent trials.

one hard instance constructed in the proof of Theorem 3.4 (Appendix E.2) for the simulation under  $f$ -divergence with  $f(x) = \alpha(x - 1)^2/2$ .

Recall that the dependency on  $\epsilon$  in all sample complexity bounds above is  $\tilde{\Theta}(\epsilon^{-1})$ , and thus both  $\text{SubOpt}_{\text{RKL}}$  and  $\text{SubOpt}_{f\text{div}}$  should be roughly proportional to  $n^{-1}$  as a function of the sample size  $n$ , which can be verified from the linear regression between  $\log_2 n$  and  $\log_2 \text{SubOpt}$ ; i.e., the estimated slope should be approximately  $-1$ . Therefore, the two fitted rates in Figure 1(a) indicates that KL-PCB indeed achieves the near-optimal statistical rate  $n^{-1}$  under different  $\pi^{\text{ref}}$ 's and the counterparts in the LHS of Figure 1(b) indicates the near-optimality of  $f$ -CB empirically. The contrast between Figure 1(a) and the LHS of Figure 1(b) also corroborates that the sample complexity against the KL-regularized objective positively depend on the concentrability, while that against the  $\chi^2$ -divergence-regularized objective does not vary with the coverage condition of  $\pi^{\text{ref}}$ . Moreover, on top of the hard instance for  $f$ -divergence, we further set  $\alpha = 2^{15}/n$  to numerically examine the scaling of  $\text{SubOpt}_{f\text{div}}$  w.r.t. the strong convexity modulus  $\alpha$ . As shown on the RHS of Figure 1(b),  $\text{SubOpt}_{f\text{div}}$  remains stable as  $n$  goes up given  $n\alpha \equiv 2^{15}$ ; therefore, Figure 1(b) also empirically verified that  $\text{SubOpt}_{f\text{div}}$  is inversely proportional to  $\alpha$ .

## 4.2 SIMULATION ON LINEAR BANDITS

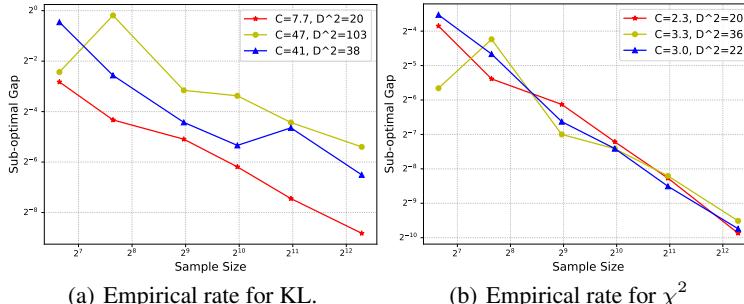


Figure 2: The empirical relation between  $\log_2 n$  and  $\log_2 \text{SubOpt}$  for linear bandits. In the legend, we denote  $C^{\pi^*}$  (resp.  $D_{\pi^*}^2$ ) by  $C$  (resp.  $D^2$ ).

We simulate a linear bandit as follows. The constructions of the feature map  $\phi$ , ground-truth parameter  $\theta^*$  and the induced reward are detailed in Appendix C.1. The behavior policy is constructed as  $\pi^{\text{ref}} = \beta \text{Unif}(\mathcal{A}) + (1 - \beta) \text{Unif}(\mathcal{A}_k)$ , where  $\mathcal{A}_k \subset \mathcal{A}$  be the subset such that  $\mathcal{A}_k$  consists of the  $k$  arms with the lowest expected reward. We consider three different behavior policies,

$(\beta, k) \in \{(1, \cdot), (0.1, 4), (0.05, 20)\}$ , which induces various  $C^{\pi^*}$  and  $D_{\pi^*}^2$  and thus enables showing the influence of coverage under different regularization. The results are compiled in Figure 2. Specifically, for results under KL-regularization depicted in Figure 2(a), we see that as the coverage coefficients  $C^{\pi^*}$  and  $D_{\pi^*}^2$  vary, there is a consistent sub-optimality gap margin between these instances. On the other hand, Figure 2(b) shows that the sub-optimality gaps under different instances (with distinct coverage coefficient) are very close for sufficiently large sample sizes. These results corroborate our theoretical finding that the sample complexity w.r.t. KL-regularized objectives is concentrability-dependent but that w.r.t  $f$ -divergence ones is not (for strongly convex  $f$ ).

### 4.3 REAL-WORLD EXPERIMENTS

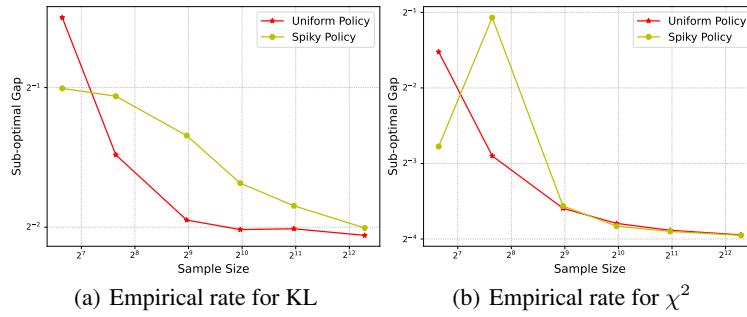


Figure 3: The empirical relation between  $\log_2 n$  and  $\log_2 \text{SubOpt}$  on MNIST dataset.

We further verify our theory on a vision dataset, MNIST (LeCun, 1998). The feature map  $\phi(\cdot, \cdot)$  construction is detailed in Appendix C.2. We consider two reference policies, a uniform policy  $\text{Unif}(\mathcal{A})$  and a spiky policy  $0.5\text{Unif}(\mathcal{A}) + 0.5\text{Dirac}(\{0\})$  to obtain instances with different concentrability coefficients. Figure 3 exhibits the SubOpt curves, which show that under KL-regularization, when sample size is not large enough, there exists a considerable gap between instances with different behavior policy, but the gap is vanishing as the sample size increases. On the other hand, as for  $\chi^2$ -divergence regularization, such a gap vanishes quickly when the sample size becomes moderately large and the sub-optimal gap remains similar for larger sample sizes. These results are consistent with the simulation in Section 4 and our theoretical findings.

## 5 CONCLUSION AND FUTURE WORK

In this work, we take the first step towards fully understanding the statistical efficiency *with respect to  $f$ -divergence-regularized objectives* of offline policy learning by sharp analyses for two empirically relevant subclasses. (1) We are the first to show that single-policy concentrability is nearly the right coverage condition for reverse KL to achieve the fast  $\tilde{\Theta}(\epsilon^{-1})$  sample complexity. The novel techniques in algorithm analysis leverages the curvature of KL-regularized objectives and integrates pessimism with a newly identified moment-based observation, enabling a neat refinement of a mean-value-type argument to the extreme; which are decoupled from tricky algorithmic tweaks, and thus might be of independent interest. (2) If strong convexity is further imposed on  $f$ , our fast  $\tilde{\Theta}(\epsilon^{-1})$  sample complexity is provably free of any coverage dependency. Unlike those for KL, the upper bound arguments for strongly convex  $f$  do not rely on specific closed-form solutions of the regularized objective maximizer.

All techniques in this work can be generalized beyond vanilla absolute reward feedback, as certified by CDBs, which is detailed in Appendix F under a slightly different notion of  $D^2$  tailored for pairwise comparison feedback. However, for reverse-KL regularization, the  $D_{\pi^*}^2$  in the upper bound and the  $C^{\pi^*}$  in the lower bound still does not perfectly match. Also, for general  $f$ -divergence other than reverse-KL, our analyses require  $f$  to be twice-continuously differentiable and strongly convex. Fully closing the gap under reverse-KL regularization and extending the analysis to general  $f$ -divergences are interesting directions for future work.

540 THE USE OF LARGE LANGUAGE MODELS (LLMs)  
541542 We use LLMs as a tool to refine our writing and correct grammatical errors.  
543544 REFERENCES  
545

546 Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the  
547 impact of entropy on policy optimization. In *International conference on machine learning*, pp.  
548 151–160. PMLR, 2019.

549 Gholamali Aminian, Amir R Asadi, Idan Shenveld, and Youssef Mroueh. Theoretical analysis of  
550 kl-regularized rlhf with multiple reference models. *arXiv preprint arXiv:2502.01203*, 2025.

551 Philip Amortila, Dylan J Foster, and Akshay Krishnamurthy. Scalable online exploration via cov-  
552 erability. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 1392–  
553 1455, 2024.

554 Michéle Basseville. Divergence measures for statistical data processing—an annotated bibliography.  
555 *Signal Processing*, 93(4):621–633, 2013.

556 Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method  
557 of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

558 Lev M Bregman. The relaxation method of finding the common point of convex sets and its applica-  
559 tion to the solution of problems in convex programming. *USSR computational mathematics and*  
560 *mathematical physics*, 7(3):200–217, 1967.

561 Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy opti-  
562 mization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

563 Fan Chen, Dylan J Foster, Yanjun Han, Jian Qian, Alexander Rakhlin, and Yunbei Xu. Assouad,  
564 fano, and le cam with interaction: A unifying lower bound framework and characterization for  
565 bandit learnability. *Advances in Neural Information Processing Systems*, 37:75585–75641, 2024.

566 Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning.  
567 In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

568 Imre Csiszár. On information-type measure of difference of probability distributions and indirect  
569 observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

570 Qiwei Di, Heyang Zhao, Jiafan He, and Quanquan Gu. Pessimistic nonlinear least-squares value  
571 iteration for offline reinforcement learning. In *The Twelfth International Conference on Learning*  
572 *Representations*, 2024.

573 Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Con-  
574 textual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

575 Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate  
576 policy and value iteration. *Advances in neural information processing systems*, 23, 2010.

577 Dylan J Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive deci-  
578 sion making. *arXiv preprint arXiv:2312.16730*, 2023.

579 Dylan J Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient  
580 reinforcement learning? the computational role of the base model in exploration. *arXiv preprint*  
581 *arXiv:2503.07453*, 2025.

582 Germano Gabbianelli, Gergely Neu, and Matteo Papini. Importance-weighted offline learning done  
583 right. In *International Conference on Algorithmic Learning Theory*, pp. 614–634. PMLR, 2024.

584 Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision  
585 processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

586 Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):  
587 504–522, 1952.

594 Dhawal Gupta, Christoph Dann, and Alekh Agarwal. P3o: Pessimistic preference-based policy  
 595 optimization for robust alignment from preferences. In *Adaptive Foundation Models: Evolving*  
 596 *AI for Personalized and Efficient Learning*, 2024.

597

598 Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy  
 599 maximum entropy deep reinforcement learning with a stochastic actor. In *International conference*  
 600 *on machine learning*, pp. 1861–1870. PMLR, 2018.

601 Jiafan He, Dongruo Zhou, and Quanquan Gu. Near-optimal policy optimization algorithms for  
 602 learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence*  
 603 *and Statistics*, pp. 4259–4280. PMLR, 2022.

604 Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster.  
 605 Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. In  
 606 *Forty-second International Conference on Machine Learning*, 2025a.

607

608 Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D Lee, Wen Sun, Akshay Krishnamurthy, and  
 609 Dylan J Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization  
 610 via chi-squared preference optimization. In *The Thirteenth International Conference on*  
 611 *Learning Representations*, 2025b.

612 Kaixuan Ji, Qingyue Zhao, Jiafan He, Weitong Zhang, and Quanquan Gu. Horizon-free reinforce-  
 613 ment learning in adversarial linear mixture mdps. *arXiv preprint arXiv:2305.08359*, 2023.

614 Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and  
 615 guarantees. *Statistical Science*, 2024.

616

617 Jiantao Jiao, Thomas A Courtade, Albert No, Kartik Venkat, and Tsachy Weissman. Information  
 618 measures: the curious case of the binary alphabet. *IEEE Transactions on Information Theory*, 60  
 619 (12):7616–7626, 2014.

620

621 Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement  
 622 learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143.  
 623 PMLR, 2020.

624

625 Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In  
 626 *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

627

628 Tadashi Kozuno, Wenhao Yang, Nino Vieillard, Toshinori Kitamura, Yunhao Tang, Jincheng Mei,  
 629 Pierre Ménard, Mohammad Gheshlaghi Azar, Michal Valko, Rémi Munos, et al. Kl-entropy-  
 630 regularized rl with a generative model is minimax optimal. *arXiv preprint arXiv:2205.14211*,  
 631 2022.

632

633 Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and  
 634 improved mitigation for reward hacking. In *The Thirteenth International Conference on Learning*  
 635 *Representations*, 2025.

636

637 Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

638

639 Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*,  
 640 pp. 38–53, 1973.

641

642 Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.

643

644 Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review.  
 645 *arXiv preprint arXiv:1805.00909*, 2018.

646

647 Sergey Levine and Vladlen Koltun. Guided policy search. In *International conference on machine*  
 648 *learning*, pp. 1–9. PMLR, 2013.

649

650 Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuo-  
 651 motor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

652

653 Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tuto-  
 654 rial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

648 Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of  
 649 model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.  
 650

651 Gene Li, Cong Ma, and Nati Srebro. Pessimism for offline linear contextual bandits using  $\ell_p$  confi-  
 652 dence sets. *Advances in Neural Information Processing Systems*, 35:20974–20987, 2022.

653 Guanlin Liu, Kaixuan Ji, Renjie Zheng, Zheng Wu, Chen Dun, Quanquan Gu, and Lin Yan. En-  
 654 hancing multi-step reasoning abilities of language models through direct q-function optimization.  
 655 *arXiv preprint arXiv:2410.09302*, 2024.

656

657 Zhuang Liu, Xuanlin Li, Bingyi Kang, and Trevor Darrell. Regularization matters in policy opti-  
 658 mization. *arXiv preprint arXiv:1910.09191*, 2019.

659 Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In *International Confer-  
 660 ence on Machine Learning*, pp. 4125–4133. PMLR, 2019.

661

662 Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in  
 663 applied probability*, 29(2):429–443, 1997.

664 Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland,  
 665 Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash  
 666 learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

667

668 Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov  
 669 decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

670 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong  
 671 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to fol-  
 672 low instructions with human feedback. *Advances in neural information processing systems*, 35:  
 673 27730–27744, 2022.

674

675 Jean-Paul Penot. Sub-hessians, super-hessians and conjugation. *Nonlinear Analysis: Theory, Meth-  
 676 ods & Applications*, 23(6):689–702, 1994.

677

678 Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge univer-  
 679 sity press, 2025.

680

681 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea  
 682 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances  
 683 in Neural Information Processing Systems*, 36, 2023.

684

685 Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline rein-  
 686 forcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information  
 687 Processing Systems*, 34:11702–11716, 2021.

688

689 Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley  
 690 symposium on mathematical statistics and probability, volume 1: contributions to the theory of  
 691 statistics*, volume 4, pp. 547–562. University of California Press, 1961.

692

693 Pierre Harvey Richemond, Yunhao Tang, Daniel Guo, Daniele Calandriello, Mohammad Gheshlaghi  
 694 Azar, Rafael Rafailov, Bernardo Avila Pires, Eugene Tarassov, Lucas Spangher, Will Ellsworth,  
 695 et al. Offline regularised reinforcement learning for large language models alignment. *arXiv  
 696 preprint arXiv:2405.19107*, 2024.

697

698 Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. Pac-bayesian offline contextual bandits with  
 699 guarantees. In *International Conference on Machine Learning*, pp. 29777–29799. PMLR, 2023.

700

701 Otmane Sakhi, Imad Aouali, Pierre Alquier, and Nicolas Chopin. Logarithmic smoothing for pes-  
 702 simistic off-policy evaluation, selection and learning. *Advances in Neural Information Processing  
 703 Systems*, 37:80706–80755, 2024.

704

705 Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline  
 706 reinforcement learning: Towards optimal sample complexity. In *International conference on  
 707 machine learning*, pp. 19967–20025. PMLR, 2022.

702 Yuda Song, Gokul Swamy, Aarti Singh, Drew Bagnell, and Wen Sun. The importance of online data:  
 703 Understanding preference fine-tuning via coverage. In *The Thirty-eighth Annual Conference on*  
 704 *Neural Information Processing Systems*, 2024.

705 Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient meth-  
 706 ods for reinforcement learning with function approximation. *Advances in neural information*  
 707 *processing systems*, 12, 1999.

708 Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from  
 709 logged bandit feedback. In *International conference on machine learning*, pp. 814–823. PMLR,  
 710 2015.

711 Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under  
 712 partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

713 Rom Rubenovich Varshamov. Estimate of the number of signals in error correcting codes. *Doklady*  
 714 *Akad. Nauk, SSSR*, 117:739–741, 1957.

715 Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist.  
 716 Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in*  
 717 *Neural Information Processing Systems*, 33:12163–12174, 2020.

718 Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cam-  
 719 bridge university press, 2019.

720 Zhiyong Wang, Dongruo Zhou, John Lui, and Wen Sun. Model-based rl as a minimalist approach  
 721 to horizon-free and second-order bounds. *arXiv preprint arXiv:2408.08994*, 2024.

722 Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforce-  
 723 ment learning. *Machine learning*, 8:229–256, 1992.

724 Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play  
 725 preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024.

726 Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent  
 727 pessimism for offline reinforcement learning. *Advances in neural information processing systems*,  
 728 34:6683–6694, 2021a.

729 Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridg-  
 730 ing sample-efficient offline and online reinforcement learning. *Advances in neural information*  
 731 *processing systems*, 34:27395–27407, 2021b.

732 Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and  
 733 Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit  $q^*$ -approximation  
 734 for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.

735 Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly mini-  
 736 max optimal offline reinforcement learning with linear function approximation: Single-agent mdp  
 737 and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

738 Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang.  
 739 Iterative preference learning from human feedback: Bridging theory and practice for rlhf under  
 740 kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

741 Yinglun Xu, Hangoo Kang, Tarun Suresh, Yuxuan Wan, and Gagandeep Singh. Learning a pes-  
 742 simistic reward model in rlhf. *arXiv preprint arXiv:2505.20556*, 2025.

743 Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. A theoretical analysis of  
 744 nash learning from human feedback under general kl-regularized preference. *arXiv preprint*  
 745 *arXiv:2402.07314*, 2024.

746 Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability*  
 747 *and statistics*, pp. 423–435. Springer, 1997.

756 Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits  
 757 problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.  
 758

759 Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement  
 760 learning with realizability and single-policy concentrability. In *Conference on Learning Theory*,  
 761 pp. 2730–2775. PMLR, 2022.

762 Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline  
 763 preference-based reinforcement learning. *arXiv preprint arXiv:2305.14816*, 2023.

764 Tong Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press,  
 765 2023. doi: 10.1017/9781009093057.

766 Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. General preference modeling  
 767 with preference representations for aligning language models. *arXiv preprint arXiv:2410.02197*,  
 768 2024.

769 Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized con-  
 770 textual bandits and rlhf. *arXiv preprint arXiv:2411.04625*, 2024.

771 Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online  
 772 kl-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460*, 2025.

773 Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration.  
 774 In *International conference on machine learning*, pp. 11492–11502. PMLR, 2020.

775 Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient.  
 776 *arXiv preprint arXiv:1803.06573*, 2018.

777 Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feed-  
 778 back from pairwise or k-wise comparisons. In *International Conference on Machine Learning*,  
 779 pp. 43037–43067. PMLR, 2023.

780 Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse  
 781 reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

782 Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence  
 783 bound for the k-armed dueling bandit problem. In *International conference on machine learning*,  
 784 pp. 10–18. PMLR, 2014.

## 785 CONTENTS

792 <b>1</b>	<b>Introduction</b>	<b>1</b>
793	1.1 Contributions . . . . .	2
794	1.2 Key Related Work . . . . .	3
795		
796		
797 <b>2</b>	<b>KL-regularized Contextual Bandits</b>	<b>4</b>
798	2.1 Problem Setup . . . . .	4
799	2.2 Algorithm . . . . .	5
800		
801		
802	2.3 Theoretical Results . . . . .	6
803	2.4 Proof Overview of Theorem 2.10 . . . . .	6
804		
805 <b>3</b>	<b><math>f</math>-divergence-regularized Contextual Bandits</b>	<b>7</b>
806	3.1 Problem Setup . . . . .	7
807	3.2 Algorithm and Main Results . . . . .	7
808		
809	3.3 Proof Overview of Theorem 3.2 . . . . .	8

810	<b>4 Experiments</b>	<b>8</b>
811	4.1 Tabular Setting . . . . .	8
812	4.2 Simulation on Linear Bandits . . . . .	9
813	4.3 Real-World Experiments . . . . .	10
814		
815		
816		
817	<b>5 Conclusion and Future Work</b>	<b>10</b>
818		
819		
820	<b>A Additional Review of Existing Results</b>	<b>17</b>
821	A.1 Previous Attempts on Understanding KL-regularized RL . . . . .	17
822		
823		
824	<b>B Additional Discussion of Relation between Coverage Measures</b>	<b>18</b>
825		
826	<b>C Experimental Details</b>	<b>19</b>
827	C.1 Linear Bandits . . . . .	19
828	C.2 Real-Wold Experiments . . . . .	19
829		
830		
831		
832	<b>D Missing Proofs from Section 2</b>	<b>19</b>
833	D.1 Proof of Lemma 2.9 . . . . .	19
834	D.2 Proof of Lemma 2.14 . . . . .	20
835	D.3 Proof of Lemma 2.15 . . . . .	21
836	D.4 Proof of Theorem 2.10 . . . . .	21
837	D.5 Proof of Theorem 2.11 . . . . .	22
838		
839		
840		
841	<b>E Missing Proof from Section 3</b>	<b>24</b>
842	E.1 Proof of Theorem 3.2 . . . . .	24
843	E.2 Proof of Theorem 3.4 . . . . .	26
844		
845		
846		
847	<b>F Generalization to contextual dueling bandits</b>	<b>27</b>
848	F.1 Problem setup . . . . .	28
849	F.2 Algorithms and Results . . . . .	28
850	F.2.1 Algorithms for KL-regularized contextual dueling bandits . . . . .	28
851	F.2.2 Algorithm and Results for $f$ -divergence Regularized CDBs . . . . .	29
852		
853		
854		
855	<b>G Missing proof from Appendix F</b>	<b>30</b>
856	G.1 Proof of Theorem F.4 . . . . .	30
857	G.2 Proof of Theorem F.6 . . . . .	32
858	G.3 Proof of Theorem F.8 . . . . .	33
859	G.4 Proof of Theorem F.9 . . . . .	34
860		
861		
862		
863	<b>H Auxiliary Lemmas</b>	<b>35</b>

## 864 A ADDITIONAL REVIEW OF EXISTING RESULTS

866 **Additional notations.** Besides the notation introduced in Section 1, we will use the following  
 867 notations in Appendix. We denote  $[N] := \{1, \dots, N\}$  for any positive integer  $N$ . Boldfaced lower  
 868 case (resp. upper case) letters are reserved for vectors (resp. matrices). Given a positive definite  
 869  $\Sigma \in \mathbb{R}^{d \times d}$  and  $\mathbf{x} \in \mathbb{R}^d$ , we denote the vector's Euclidean norm by  $\|\mathbf{x}\|_2$  and define  $\|\mathbf{x}\|_\Sigma =$   
 870  $\sqrt{\mathbf{x}^\top \Sigma \mathbf{x}}$ . We use  $\text{Bern}(p)$  to denote Bernoulli distribution with expectation  $p$  and  $\text{Unif}(\mathcal{X})$  for the  
 871 uniform distribution on finite set  $\mathcal{X}$ . For  $x \in \mathbb{R}^{|\mathcal{A}|}$ , we denote  $\|x\|_1 = \sum_{a \in \mathcal{A}} |x_a|$ . We also denote  
 872  $x_n = \Omega(y_n)$  by  $x_n \gtrsim y_n$  in Appendix. We use  $d_H$  for Hamming distance.

### 873 A.1 PREVIOUS ATTEMPTS ON UNDERSTANDING KL-REGULARIZED RL

875 There has been a surge of interest in understanding the principle behind KL-regularized RL. Ahmed  
 876 et al. (2019); Liu et al. (2019) studied by ablation the effect of entropy regularization on the stability  
 877 of policy improvement in policy optimization, the regret of which has been rigorously settled under  
 878 the classic online mirror descent framework (Cai et al., 2020; He et al., 2022; Ji et al., 2023). Neu  
 879 et al. (2017) unified popular KL-regularized policy optimization algorithms under a convex opti-  
 880 mization framework, but the interplay with the data was left untouched. A series of work (Geist  
 881 et al., 2019; Vieillard et al., 2020; Kozuno et al., 2022) then analyzed the sample complexity of  
 882 algorithms using KL/entropy-type proximal terms with respect to the previous iteration or/and en-  
 883 tropy regularizer with improved dependence on the effective horizon in discounted Markov decision  
 884 processes. However, the performance metric in these studies is still the unregularized reward max-  
 885 imization objective, under which the sample complexity for finding an  $\epsilon$ -optimal policy is at least  
 886 equal to the statistical limit  $\Omega(\epsilon^{-2})$ .

887 **Convergence under regularized objectives.** Several recent studies (Xie et al., 2024; Xiong et al.,  
 888 2024; Zhao et al., 2024; 2025; Foster et al., 2025) switched the focus to analyzing the sub-optimality  
 889 guarantee with respect to the regularized objective (1.1). In particular, Xie et al. (2024) studied  
 890 token-level Markov decision processes (MDPs) and proposed a KL-regularized RL algorithm  
 891 named XPO, which achieves  $\tilde{O}(\epsilon^{-2})$  sample complexity under their notion of all-policy concentrability.  
 892 Xiong et al. (2024) proposed an Offline GSHF algorithm via the principle of *pessimism in the*  
 893 *face of uncertainty*, and proved  $\tilde{O}(\epsilon^{-2})$  sample complexity under single-policy concentrability (See  
 894 Section 2.1 for detailed definitions of concentrability). On the other hand, the sharp analysis in Zhao  
 895 et al. (2024) yields the optimal sample complexity  $\tilde{O}(\epsilon^{-1})$ , but requires all-policy concentrability  
 896 (Zhao et al., 2024, Definition 2.6), i.e., the behavior policy  $\pi^{\text{ref}}$  is required to cover the entire function  
 897 class for all possible policies. Zhao et al. (2025) considered the online episodic MDP setting,  
 898 which inherently does not need any notion of data coverage and thus their results are not directly  
 899 adaptable to our offline setting. Foster et al. (2025) considered an interesting hybrid setting in which  
 900 the  $n$  state-action pairs are still from the offline dataset but  $\Omega(n)$  online reward queries and policy  
 901 switches are allowed; in contrast, in our setting, all reward signals are obtained in a purely offline  
 902 manner.

903 **Previous analyses and results in detail.** Here, we briefly discuss the direct adaptation of previous  
 904 sample complexity analysis and results (with respect to KL-regularized objectives) to our setting and  
 905 demonstrate the reason why theirs cannot imply an  $\tilde{O}(\epsilon^{-1})$  sample complexity without all-policy  
 906 concentrability. In previous analysis of pessimism for unregularized objectives (Jin et al., 2021;  
 907 Xiong et al., 2022), the sub-optimality gap is decomposed via the performance difference lemma as  
 908 follows

$$\begin{aligned}
 910 J(\pi^*) - J(\hat{\pi}) &= \mathbb{E}_{a \sim \pi^*}[g^*(a)] - \mathbb{E}_{a \sim \hat{\pi}}[g^*(a)] - \eta^{-1} \text{KL}(\pi^* \parallel \pi^{\text{ref}}) + \eta^{-1} \text{KL}(\hat{\pi} \parallel \pi^{\text{ref}}) \\
 911 &\leq \mathbb{E}_{a \sim \pi^*}[g^*(a)] - \mathbb{E}_{a \sim \hat{\pi}}[\hat{g}(a)] - \eta^{-1} \text{KL}(\pi^* \parallel \pi^{\text{ref}}) + \eta^{-1} \text{KL}(\hat{\pi} \parallel \pi^{\text{ref}}) \\
 912 &\leq \mathbb{E}_{a \sim \pi^*}[g^*(a)] - \mathbb{E}_{a \sim \pi^*}[\hat{g}(a)] - \eta^{-1} \text{KL}(\pi^* \parallel \pi^{\text{ref}}) + \eta^{-1} \text{KL}(\pi^* \parallel \pi^{\text{ref}}) \\
 913 &= \mathbb{E}_{a \sim \pi^*}[g^*(a) - \hat{g}(a)],
 914
 \end{aligned}$$

915 where the first inequality holds due to pessimism and last inequality holds due to  $\hat{\pi}$  is optimal for  
 916  $\hat{g}$ . Notably, the KL-regularization term is canceled out in the analysis, leading to a loose sample  
 917 complexity  $\tilde{O}(\epsilon^{-2})$  since the curvature of KL-divergence is not exploited. Specifically, under linear

918 function approximation, this performance gap, obtained by Xiong et al. (2024) becomes  
919

$$920 \quad J(\pi^*) - J(\pi) \leq \|\mathbb{E}_{\rho \times \pi^*}[\phi(s, a)] - \nu\|_{\Sigma_{\text{off}}^{-1}} =: \text{RHS},$$

921 where  $\nu$  is the reference vector,  $\phi(s, a) \in \mathbb{R}^d$  is the feature map, and  $\Sigma_{\text{off}} = \sum_{i=1}^n \phi(s_i, a_i) \phi(s_i, a_i)^\top$  is the sample covariance matrix. However, we can show that RHS can  
922 be bounded from below by  
923

$$924 \quad \|\mathbb{E}_{(s, a) \sim \rho \times \pi^*}[\phi(s, a)] - \nu\| \sqrt{\lambda_{\min}(\Sigma_{\text{off}}^{-1})} = \|\mathbb{E}_{(s, a) \sim \rho \times \pi^*}[\phi(s, a)] - \nu\| \lambda_{\max}(\Sigma_{\text{off}})^{-1/2} \\ 925 \quad \geq \|\mathbb{E}_{(s, a) \sim \rho \times \pi^*}[\phi(s, a)] - \nu\| \text{tr}(\Sigma_{\text{off}})^{-1/2} \\ 926 \quad = \|\mathbb{E}_{(s, a) \sim \rho \times \pi^*}[\phi(s, a)] - \nu\| \left( \sum_{i=1}^n \|\phi(s_i, a_i)\|_2^2 \right)^{-1/2} \\ 927 \quad = \Omega(n^{-1/2}),$$

928 where  $\lambda_{\min}$  and  $\lambda_{\max}$  is the minimum and maximum eigenvalue of a matrix, the first inequality  
929 holds due to the fact that  $\mathbf{x}^\top \Sigma \mathbf{x} \geq \|\mathbf{x}\|_2^2 \lambda_{\min}(\Sigma)$  and the second inequality holds due to  $\lambda_{\max}(\Sigma) \leq$   
930  $\text{tr}(\Sigma)$ . Zhao et al. (2024) proposed a two-stage learning algorithm and obtained an  $\tilde{O}(\epsilon^{-1})$  sample  
931 complexity for online KL-regularized bandits. The algorithm can be adopted to offline learning by  
932 removing the second stage<sup>5</sup> and treat the samples from first stage as the offline dataset. An analogous  
933 analysis gives a sample complexity of  $\tilde{O}(D^2 \epsilon^{-1})$ , where  $D^2$  is the all-policy concentrability.  
934

## 935 B ADDITIONAL DISCUSSION OF RELATION BETWEEN COVERAGE 936 MEASURES

937 In this section, we provide more illustrations on the relation between two coverage measures,  $D_{\pi^*}^2$   
938 and  $C^{\pi^*}$ . In particular, we provide two cases under linear function approximation, on one of which  
939  $D_{\pi^*}^2 = \Theta(dC^{\pi^*})$  and on the other we have  $D_{\pi^*}^2 \ll C^{\pi^*}$ , where  $d$  is the dimension of the function  
940 class. We summarized them as two propositions.  
941

942 **Proposition B.1.** There exist a KL-regularized linear bandit instance, such that  $D_{\pi^*}^2 = \Theta(dC^{\pi^*})$ .  
943

944 *Proof.* We construct the instance as follows. Let  $d = 2A + 1$  be some odd number and consider an  
945  $2A + 1$ -armed bandit, such that the feature vector of the  $i$ -th arm,  $\phi(a_i) = \mathbf{e}_i \in \mathbb{R}^d$ , which has 1 on  
946 its  $i$ -th entry and 0 on all other entries. The reference policy  $\pi^{\text{ref}}(a_i) = (2AC)^{-1}$  for  $i \in [2A]$  and  
947  $\pi^{\text{ref}}(a_{2A+1}) = (C - 1)/C$ , where  $2C - 1 = e^\eta$ . The ground truth reward function  $\theta^* = \sum_{i \leq A} \mathbf{e}_i$   
948 and the function class is given by all  $\|\theta\|_\infty \leq 1$ . By construction, we know that  $\pi^*(a_i) \geq \pi^{\text{ref}}(a_i)$   
949 if and only if  $i \in [A]$  and its closed form is given by  
950

$$951 \quad \pi^*(a_i) = \frac{1}{A} \frac{e^\eta}{e^\eta + 2C - 1} = \frac{1}{2A},$$

952 which gives  $C^{\pi^*} = C$ . Now we compute the  $D_{\pi^*}^2$  of this instance. For all  $i \in [A]$ , we know that  
953

$$954 \quad D^2(a_i) = \sup_{\|\theta\|_\infty \leq 2} \frac{\langle \theta, \mathbf{e}_i \rangle^2}{\mathbb{E}_{\pi^{\text{ref}}} \langle \theta, \mathbf{e}_i \rangle^2} = 2CA = \Theta(Cd),$$

955 where the second equation holds with  $\theta = \mathbf{e}_i$ . Taking expectation over  $\pi^*$ , we have  
956

$$957 \quad D_{\pi^*}^2 \geq \sum_{i \in [A]} D^2(a_i) = \Theta(C^{\pi^*} d),$$

958 which concludes the proof. □  
959

960 The following proposition provides another instance on which  $D_{\pi^*}^2 \ll C^{\pi^*}$ .  
961

962 **Proposition B.2.** For any  $C \geq 2$ , there exists a KL-regularized linear bandit instance, such that  
963  $C^{\pi^*} = C/2$  and  $D_{\pi^*}^2 = \Theta(1)$ .  
964

965 <sup>5</sup>This can be done by setting the  $n$  in their paper to 0.  
966

972 *Proof.* We consider the function class of  $\theta \in \mathbb{R}^2$  and  $\|\theta\| \leq \sqrt{2}$ . The instance consists of three arms,  
 973 where  $\phi(a_1) = (1, 0)$ ,  $\phi(a_2) = (0, 1)$ , and  $\phi(a_3) = (1, 1)$ . The ground truth parameter  $\theta^* = (1, 1)$ .  
 974 The reference policy is given by  $\pi^{\text{ref}}(a_1) = \pi^{\text{ref}}(a_2) = 1/2 - 1/2C$  and  $\pi^{\text{ref}}(a_3) = 1/C$ , where  
 975  $C - 1 = e^\eta$ . A direct computation yields that

$$\pi^*(a_3) = \frac{e^\eta}{e^\eta + C - 1}, \quad \Rightarrow \quad C^{\pi^*} = C \frac{e^\eta}{e^\eta + C - 1} = \frac{C}{2}.$$

976 On the other hand, we know that for  $i = 1, 2$ , we have  $D^2(a_i) \leq \pi^{\text{ref}}(a_i)^{-1} \leq 4$ . As for  $a_3$ ,  
 977 since we have  $\langle \theta, \phi(a_3) \rangle^2 = \langle \theta, \phi(a_1) + \phi(a_2) \rangle^2 \leq 2 \langle \theta, \phi(a_1) \rangle^2 + 2 \langle \theta, \phi(a_2) \rangle$ , which gives  
 978 that  $D^2(a_3) \leq 2D^2(a_1) + 2D^2(a_2) \leq 16$ . Therefore, taking expectation over  $\pi^*$ , we know that  
 979  $D_{\pi^*}^2 \leq 12$  which is a constant.  $\square$

## 983 C EXPERIMENTAL DETAILS

### 984 C.1 LINEAR BANDITS

985 The linear bandit instance used for Figure 2 has  $d = 20$  and  $|\mathcal{A}| = 100$ . For each arm  $a \in \mathcal{A}$ ,  
 986 we randomly generate its feature vector  $\phi(a) \in \mathbb{R}^d$  such that  $\|\phi(a)\| = 1$ . We then randomly  
 987 sample the model parameter  $\theta^* \in \mathbb{R}^d$  such that  $\|\theta^*\| = 1$  and the expected reward is obtained via  
 988  $r(a) = \langle \theta^*, \phi(a) \rangle$ .

### 989 C.2 REAL-WOLD EXPERIMENTS

990 MNIST consists of 60000 figures, each of which is of  $28 \times 28$  pixels and consists of a handwritten  
 991 digit in  $\{0, \dots, 9\}$ . Here, we consider each image as a context and  $\mathcal{A} = \{0, \dots, 9\}$  for each  
 992 context. To obtain the feature  $\phi(s, a)$ , we first use the hidden representation of a classifier to embed  
 993 each image as a vector in  $\mathbb{R}^{10}$ . We then follow the approach in Zhou et al. (2020) to obtain the  
 994 feature of each context-action pair by having  $\phi(s, a) = \mathbf{x} \otimes \mathbf{e}_{a+1} \in \mathbb{R}^{100}$ , where  $\mathbf{x}$  is the output of  
 995 image encoder and  $\otimes$  stands for tensor product.

## 996 D MISSING PROOFS FROM SECTION 2

### 1000 D.1 PROOF OF LEMMA 2.9

1001 We first provide the following lemmas of concentration.

1002 **Lemma D.1** (Zhao et al. 2024, Lemma C.1). For any policy  $\pi$  and state-action pairs  $\{(s_i, a_i)\}_{i=1}^m$   
 1003 generated i.i.d. from  $\rho \times \pi$ , and  $\epsilon_c < 1$ , with probability at least  $1 - \delta$ , for any  $g_1$  and  $g_2$  we have

$$1004 \mathbb{E}_{\rho \times \pi} [(g_1(s, a) - g_2(s, a))^2] \leq \frac{2}{n} \sum_{i=1}^n (g_1(s_i, a_i) - g_2(s_i, a_i))^2 + \frac{32}{3n} \log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 10\epsilon_c,$$

1005 where  $\mathcal{N}_{\mathcal{G}}(\epsilon_c)$  is the  $\epsilon_c$ -covering number of  $\mathcal{G}$ .

1006 **Lemma D.2** (Zhao et al. 2024, Lemma C.2). For arbitrary policy  $\pi$  and dataset  $\{(s_i, a_i, r_i)\}_{i=1}^m$   
 1007 generated i.i.d., from the product of  $\pi$ ,  $\rho$  and the Bradley-Terry Model; let  $\bar{g}$  be the least square  
 1008 estimator of  $g^*$ , then for any  $0 < \epsilon_c < 1$  and  $\delta > 0$ , with probability at least  $1 - \delta$  we have

$$1009 \sum_{i=1}^n (\bar{g}(s_i, a_i) - g^*(s_i, a_i))^2 \leq 16 \log(a\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 4n\epsilon_c.$$

1010 Now we are ready to prove Lemma 2.9.

1011 *Proof of Lemma 2.9.* We have the following inequality

$$1012 \begin{aligned} 1013 (\bar{g}(s, a) - g^*(s, a))^2 &= \frac{(\bar{g}(s, a) - g^*(s, a))^2}{\mathbb{E}_{\pi^{\text{ref}}}[(\bar{g}(s, a) - g^*(s, a))^2]} \mathbb{E}_{\pi^{\text{ref}}}[(\bar{g}(s, a) - g^*(s, a))^2] \\ 1014 &\leq \sup_{g_1, g_2 \in \mathcal{G}} \frac{(g_1(s, a) - g_2(s, a))^2}{\mathbb{E}_{\pi^{\text{ref}}}[(g_1(s, a) - g_2(s, a))^2]} \mathbb{E}_{\pi^{\text{ref}}}[(\bar{g}(s, a) - g^*(s, a))^2] \\ 1015 &= D_{\mathcal{G}}^2((s, a), \pi^{\text{ref}}) \mathbb{E}_{\pi^{\text{ref}}}[(\bar{g}(s, a) - g^*(s, a))^2], \end{aligned} \tag{D.1}$$

where the inequality holds by taking supremum to  $g_1, g_2 \in \mathcal{G}$ . Now we have

$$\begin{aligned}
\mathbb{E}_{\pi^{\text{ref}}} \left[ (\bar{g}(s, a) - g^*(s, a))^2 \right] &\leq \frac{2}{n} \sum_{i=1}^n (\bar{g}(s_i, a_i) - g^*(s_i, a_i))^2 + \frac{32}{3n} \log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 10\epsilon_c \\
&\leq \frac{2}{n} [16 \log(\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 4n\epsilon_c] + \frac{32}{3n} \log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 10\epsilon_c \\
&= \frac{128}{3n} \log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 18\epsilon_c,
\end{aligned} \tag{D.2}$$

where the first inequality holds due to Lemma D.1 and second holds due to Lemma D.2. Plugging (D.2) into (D.1) and setting  $\epsilon_c = O(n^{-1})$  complete the proof.  $\square$

## D.2 PROOF OF LEMMA 2.14

This proof is extracted from the proof of Zhao et al. (2024, Theorem 3.3) and we present it here for completeness. By definition of our objective in (2.1), we have

$$\begin{aligned}
& J(\pi^*) - J(\pi_g) \\
&= \mathbb{E}_{(s,a) \sim \rho \times \pi^*} \left[ g^*(s,a) - \eta^{-1} \log \frac{\pi^*(a|s)}{\pi^{\text{ref}}(a|s)} \right] - \mathbb{E}_{(s,a) \sim \rho \times \pi_g} \left[ g^*(s,a) - \frac{1}{\eta} \log \frac{\pi_g(a|s)}{\pi^{\text{ref}}(a|s)} \right] \\
&= \frac{1}{\eta} \mathbb{E}_{(s,a) \sim \rho \times \pi^*} \left[ \log \frac{\pi^{\text{ref}}(a|s) \cdot \exp(\eta g^*(s,a))}{\pi^*(a|s)} \right] - \frac{1}{\eta} \mathbb{E}_{(s,a) \sim \rho \times \pi_g} \left[ \log \frac{\pi^{\text{ref}}(a|s) \cdot \exp(\eta g^*(s,a))}{\pi_g(a|s)} \right] \\
&= \frac{1}{\eta} \mathbb{E}_{s \sim \rho} \left[ \log Z_{g^*}(s) \right] - \frac{1}{\eta} \mathbb{E}_{s \sim \rho} \left[ \log Z_g(s) \right] - \mathbb{E}_{s \sim \rho} \left[ \sum_{a \in \mathcal{A}} \pi_g(a|s) \cdot (g^*(s,a) - f(s,a)) \right],
\end{aligned}$$

where for all  $g \in \mathcal{G}$  we define  $Z_g(\cdot)$  as follows,

$$Z_g(\cdot) := \sum_{a \in \mathcal{A}} \pi^{\text{ref}}(a|\cdot) \exp(\eta g(\cdot, a)).$$

We further denote  $\Delta(s, a) = g(s, a) - g^*(s, a)$  and  $H_s(g) = \log Z_g(s) - \eta \sum_{a \in \mathcal{A}} \pi_g(a|s) \cdot \Delta(s, a)$ . It is worth noticing that  $\eta^{-1} \mathbb{E}_{s \sim \rho} [H_s(g^*) - H_s(g)] = J(\pi^*) - J(\pi_g)$ . Now we take derivative of  $H$  with respect to  $\Delta(s, a)$ ,

$$\begin{aligned}
\frac{\partial H_s(g)}{\partial \Delta(s, a)} &= \frac{\partial}{\partial \Delta(s, a)} \left[ \log Z_g(s) - \eta \sum_{a \in \mathcal{A}} \pi_g(a|s) \cdot \Delta(s, a) \right] \\
&= \frac{1}{Z_g(s)} \cdot \pi^{\text{ref}}(a|s) \exp(\eta \cdot g(s, a)) \cdot \eta - \eta \cdot \pi_g(a|s) \\
&\quad - \eta^2 \cdot \Delta(s, a) \cdot \frac{\pi^{\text{ref}}(a|s) \cdot \exp(\eta \cdot g(s, a))}{Z_g(s)} + \eta^2 \cdot \Delta(s, a) \cdot \frac{[\pi^{\text{ref}}(a|s) \cdot \exp(\eta \cdot g(s, a))]^2}{[Z_g(s)]^2} \\
&\quad + \eta \sum_{a' \in \mathcal{A} \setminus \{a\}} \frac{\pi^{\text{ref}}(a'|s) \cdot \exp(\eta \cdot g(s, a'))}{Z_g(s)} \cdot \eta \cdot \Delta(s, a') \cdot \frac{\pi^{\text{ref}}(a|s) \cdot \exp(\eta \cdot g(s, a))}{Z_g(s)} \\
&= -\eta^2 \pi_g(a|s) \Delta(s, a) + \eta^2 [\pi_g(a|s)]^2 \cdot \Delta(s, a) + \eta^2 \sum_{a' \in \mathcal{A} \setminus \{a\}} \pi_g(a'|s) \pi_g(a|s) \Delta(s, a').
\end{aligned}$$

Therefore, by mean value theorem, there exists  $\gamma \in [0, 1]$  and  $q_\gamma = \gamma q + (1 - \gamma)q^*$  such that

$$\begin{aligned}
H_s(g) - H_s(g^*) &= -\eta^2 \gamma \sum_{a \in \mathcal{A}} \pi_{g_\gamma}(a|s) \Delta(s, a)^2 + \gamma \eta^2 \sum_{a_1 \in \mathcal{A}} \sum_{a_2 \in \mathcal{A}} \pi_{g_\gamma}(a_1|x) \pi_{g_\gamma}(a_2|x) \Delta(s, a_1) \Delta(s, a_2) \\
&= -\eta^2 \gamma \mathbb{E}_{a \sim \pi_{g_\gamma}} \left[ (g^*(s, a) - g(s, a))^2 \right] + \gamma \eta^2 \left( \mathbb{E}_{a \sim \pi_{g_\gamma}} \left[ (g^*(s, a) - g(s, a)) \right] \right)^2 \\
&\geq -\eta^2 \mathbb{E}_{a \sim \pi_{g_\gamma}} \left[ (g^*(s, a) - g(s, a))^2 \right],
\end{aligned}$$

1080 where the inequality holds by omitting the second term and  $\gamma \leq 1$ . Now taking expectation over  $\rho$ ,  
 1081 we have

$$\begin{aligned} 1083 \quad J(\pi^*) - J(\pi_g) &= \eta^{-1} \mathbb{E}_{s \sim \rho} [H_s(g^*) - H_s(g)] \\ 1084 &\leq \eta \mathbb{E}_{(s,a) \sim \rho_{g_\gamma}} [(g^*(s,a) - g(s,a))^2], \end{aligned}$$

1086 which concludes the proof.

### 1087 D.3 PROOF OF LEMMA 2.15

1089 *Proof of Lemma 2.15.* We define  $Y = -X$ . Then it suffices to show that the covariance between  $Y$   
 1090 and  $Y^2$  is

$$\begin{aligned} 1091 \quad \text{Cov}(Y, Y^2) &= \mathbb{E}[Y^3] - \mathbb{E}[Y^2]\mathbb{E}[Y] \\ 1092 &\geq (\mathbb{E}[Y^2])^{3/2} - \mathbb{E}[Y^2]\mathbb{E}[Y] \\ 1093 &= (\mathbb{E}[Y^2])(\sqrt{\mathbb{E}[Y^2]} - \mathbb{E}[Y]) \\ 1094 &\geq 0, \\ 1095 \\ 1096 \end{aligned}$$

1097 where both inequalities follow from Jensen's inequality.  $\square$

### 1099 D.4 PROOF OF THEOREM 2.10

1101 To start with, we first define the following quantities. For all  $\gamma \in [0, 1]$ , we define  $g_\gamma := \gamma \hat{g} + (1 -$   
 1102  $\gamma)g^*$  and denote

$$\begin{aligned} 1103 \quad \pi_\gamma(\cdot|s) &\propto \pi^{\text{ref}}(\cdot|s) \exp(\eta g_\gamma(s, \cdot)), \forall s \in \mathcal{S}; \\ 1104 \quad G(\gamma) &:= \mathbb{E}_{\rho \times \pi_\gamma} [(\hat{g} - g^*)^2(s, a)]. \end{aligned}$$

1107 The key to our analysis is the monotonicity of the function  $G(\gamma)$  in  $\gamma$ , which is formally stated in  
 1108 the following lemma.

1109 **Lemma D.3.** On event  $\mathcal{E}$ ,  $0 \in \text{argmax}_{\gamma \in [0,1]} G(\gamma)$ .

1112 *Proof.* For simplicity, we use  $\Delta(s, a)$  to denote  $(\hat{g} - g^*)(s, a)$  in this proof. Then we know that  
 1113  $\Delta(s, a) \leq 0$  for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$  on event  $\mathcal{E}$ . The most direct way to prove is to take derivative  
 1114 of  $G$  with respect to  $\gamma$ , which corresponds to the policy gradient (Sutton et al., 1999) of  $\pi_\gamma$  and thus  
 1115 implying a favorable structure. A direct calculation yields that

$$\begin{aligned} 1116 \quad &= \mathbb{E}_{\rho \times \pi_\gamma} [\nabla_\gamma \log \pi_\gamma(a|s) \Delta(s, a)^2] \\ 1117 &= \eta \mathbb{E}_\rho \mathbb{E}_{a \sim \pi_\gamma} [\Delta^2(s, a) (\Delta(s, a) - \mathbb{E}_{a' \sim \pi_\gamma} [\Delta(s, a')])] \\ 1118 &= \eta \mathbb{E}_\rho [\mathbb{E}_{\pi_\gamma} [\Delta^3(s, a)] - \mathbb{E}_{\pi_\gamma} [\Delta^2(s, a)] \mathbb{E}_{\pi_\gamma} [\Delta(s, a)]] \\ 1119 &\leq 0, \\ 1120 \\ 1121 \end{aligned}$$

1122 where  $\mathbb{E}_\rho$  is the shorthand of  $\mathbb{E}_{s \sim \rho}$ ,  $\mathbb{E}_{\pi_\gamma}$  is the shorthand of  $\mathbb{E}_{a \sim \pi_\gamma}$ , the first equation is derived from  
 1123 standard policy gradient and the inequality holds conditioned on the event  $\mathcal{E}(\delta)$  due to Lemma 2.15  
 1124 and Lemma 2.15.  $\square$

1126 Now we are ready to prove Theorem 2.10.

1128 *Proof of Theorem 2.10.* Following the proof of Zhao et al. (2024, Theorem 3.3), we know that there  
 1129 exists  $\bar{\gamma} \in [0, 1]$  such that

$$1131 \quad J(\pi^*) - J(\hat{\pi}) \leq \eta G(\bar{\gamma}) \leq \eta G(0), \quad (\text{D.3})$$

1132 where the first inequality holds due to Lemma 2.14 and the second inequality holds due to the event  
 1133  $\mathcal{E}$  and Lemma D.3. The term  $G(0)$  can be further bounded with the  $D^2$ -based concentrability as

1134 follows

$$\begin{aligned}
G(0) &= \eta \mathbb{E}_{(s,a) \sim \rho \times \pi^*} \left[ (\hat{g} - g^*)^2(s, a) \right] \\
&\leq 4\eta \mathbb{E}_{(s,a) \sim \rho \times \pi^*} [\Gamma_n^2(s, a)] \\
&= 4\eta \beta^2 \mathbb{E}_{(s,a) \sim \rho \times \pi^*} [D_{\mathcal{F}}^2((s, a); \pi^{\text{ref}})] \\
&= \tilde{O}(\eta D_{\pi^*}^2 n^{-1} \log_{\mathcal{G}}(\epsilon_c)),
\end{aligned} \tag{D.4}$$

1142 where the second inequality holds conditioned on  $\mathcal{E}(\delta)$  because of Lemma D.3, and the last inequality  
 1143 follows from the definition of  $\mathcal{E}(\delta)$  together with Line 2. By Lemma 2.9, we know that event  $\mathcal{E}$   
 1144 holds with probability at least  $1 - \delta$ , which finishes the proof.  $\square$

## D.5 PROOF OF THEOREM 2.11

*Proof of Theorem 2.11.* We consider the family of contextual bandits with  $S := |\mathcal{S}|$ ,  $A := |\mathcal{A}| < \infty$  and reward function in some function class  $\mathcal{G}$  composed of function  $\mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$  as follows.

$$\text{CB}_{\mathcal{G}} \coloneqq \{(\mathcal{S}, \mathcal{A}, \rho, r, \pi^{\text{ref}}, \eta) : r \in \mathcal{G}, \rho \in \Delta(\mathcal{S}), \pi^{\text{ref}} \in \Delta(\mathcal{A}|\mathcal{S})\}. \quad (\text{D.5})$$

Our goal is to prove the following statement. Fixing any  $S \geq 1$ ,  $\eta > 4 \log 2$  and  $C^* \in (2, \exp(\eta/4)]$ , then for any estimator  $\mathcal{D} \mapsto \hat{\pi} \in \Delta(\mathcal{A}|\mathcal{S})$ , for any  $n \geq 16SC^*$ , there exist some function class  $\mathcal{G}$ , such that  $\exists \text{inst} = (\mathcal{S}, \mathcal{A}, \rho, r, \pi^{\text{ref}}, \eta) \in \text{CB}_{\mathcal{G}}$  with single-policy concentrability  $C^{\pi^*} \leq C^*$ , regularization coefficient  $\eta$ ,  $|\mathcal{S}| = S = \Theta(\log |\mathcal{G}|)$ , and

$$\text{SubOpt}_{\text{BKL}}(\hat{\pi}; \text{inst}) \gtrsim \min\{\eta S C^* n^{-1}, (S C^*)^{1/2} n^{-1/2}\}. \quad (\text{D.6})$$

Since  $\log |\mathcal{G}| \geq \log \mathcal{N}_{\mathcal{G}}(\epsilon)$  for any  $\epsilon \in (0, 1)$ , equation (D.6) yields the desired bound.

We set  $S = [S]$ ,  $\mathcal{A} = \{\pm 1\}$ ,  $\rho = \text{Unif}(S)$ , and the reference policy to be

$$\forall s \in \mathcal{S}, \pi^{\text{ref}}(-1|s) = C^{-1}, \pi^{\text{ref}}(+1|s) = 1 - C^{-1};$$

where  $C \geq 1$  is a parameter to be specified later. We construct  $2^S$  Bernoulli reward functions, in particular,  $\forall \tau \in \{\pm 1\}^S$ , the mean function  $r_\tau$  of the reward (indexed by  $\tau$ ) is defined as

$$r_\tau(s, -1) = 0.5 + \tau_s \delta, r_\tau(s, +1) = 0.5 - \alpha$$

for any state  $s \in \mathcal{S}$ , where  $\alpha \in (0, 1/2)$  and  $\delta \in (0, 1/4]$  will be specified later. We omit the RKL subscript in the following argument when it is clear in context. By (2.2), the optimal policy  $\pi_\tau^*$  under  $r_\tau$  is

$$\forall s \in \mathcal{S}, \pi_\tau^*(-1|s) = \frac{\exp(\eta(\alpha + \tau_s \delta))}{\exp(\eta(\alpha + \tau_s \delta)) + C - 1}, \pi_\tau^*(+1|s) = \frac{C - 1}{\exp(\eta(\alpha + \tau_s \delta)) + C - 1}. \quad (\text{D.7})$$

Since  $C^* \leq \exp(\eta/4)$ , we assign  $C = C^*$  and  $\alpha = \eta^{-1} \log(C - 1) \Leftrightarrow C - 1 = \exp(\eta\alpha)$ , which gives

$$\forall s \in \mathcal{S}, \frac{\pi_\tau^*(-1|s)}{\pi^{\text{ref}}(-1|s)} \leq C \frac{\exp(\eta(\alpha + \tau s \delta))}{C - 1 + \exp(\eta(\alpha + \tau s \delta))} = C \frac{\exp(\eta \tau s \delta)}{1 + \exp(\eta \tau s \delta)} \leq C = C^*;$$

$$\forall s \in \mathcal{S}, \frac{\pi_\tau^*(+1|s)}{\pi^{\text{ref}}(+1|s)} = \frac{C}{C-1} \cdot \frac{1}{\exp(\eta\pi_\tau\delta) + 1} \leq C = C^*;$$

where the last inequality is due to the assumption  $C^* \geq 2$ . Therefore, we obtain

$$\max_{\tau \in \{1, \dots, S\}} C^{\pi_\tau^*} \leq C^*. \quad (\text{D.8})$$

We will abuse the notation  $\text{SubOpt}(\hat{\pi}; \tau) := \text{SubOpt}(\hat{\pi}; r_{-})$ . Since  $\rho \equiv \text{Unif}(S)$

$$\text{SubOpt}(\widehat{\pi}; \tau) = \frac{1}{S} \sum_s \text{SubOpt}_s(\widehat{\pi}; \tau), \quad (\text{D.9})$$

1188 where

$$\begin{aligned}
 \text{SubOpt}_s(\widehat{\pi}; \tau) &= \langle \pi_\tau^*(\cdot|s), r_\tau(s, \cdot) - \eta^{-1} \log \frac{\pi_\tau^*(\cdot|s)}{\pi^{\text{ref}}(\cdot|s)} \rangle - \langle \widehat{\pi}(\cdot|s), r_\tau(s, \cdot) - \eta^{-1} \log \frac{\widehat{\pi}(\cdot|s)}{\pi^{\text{ref}}(\cdot|s)} \rangle \\
 &= \frac{1}{\eta} \mathbb{E}_{a \sim \pi_\tau^*(\cdot|s)} \left[ \log \frac{\pi^{\text{ref}}(a|s) \cdot \exp(\eta r_\tau(s, a))}{\pi_\tau^*(a|s)} \right] \\
 &\quad - \frac{1}{\eta} \mathbb{E}_{a \sim \widehat{\pi}(\cdot|s)} \left[ \log \frac{\pi^{\text{ref}}(a|s) \cdot \exp(\eta r_\tau(s, a))}{\widehat{\pi}(a|s)} \right] \\
 &= \frac{1}{\eta} \mathbb{E}_{a \sim \pi_\tau^*(\cdot|s)} \left[ \log \left( \sum_{b \in \mathcal{A}} \pi^{\text{ref}}(b|s) \cdot \exp(\eta r_\tau(s, b)) \right) \right] \\
 &\quad - \frac{1}{\eta} \mathbb{E}_{a \sim \widehat{\pi}(\cdot|s)} \left[ \log \frac{\pi^{\text{ref}}(a|s) \cdot \exp(\eta r_\tau(s, a))}{\widehat{\pi}(a|s)} \right] \\
 &= \frac{1}{\eta} \mathbb{E}_{a \sim \widehat{\pi}(\cdot|s)} \left[ \log \frac{\pi^{\text{ref}}(a|s) \cdot \exp(\eta r_\tau(s, a))}{\pi_\tau^*(a|s)} - \log \frac{\pi^{\text{ref}}(a|s) \cdot \exp(\eta r_\tau(s, a))}{\widehat{\pi}(a|s)} \right] \\
 &= \eta^{-1} \mathbf{KL}(\widehat{\pi} \parallel \pi_\tau^*). \tag{D.10}
 \end{aligned}$$

1206 We write  $\tau \sim_s \tau'$  if  $\tau, \tau' \in \{\pm 1\}^{\mathcal{S}}$  differ in only the  $s$ -th coordinate and  $\tau \sim \tau'$  if  $\exists s \in \mathcal{S}, \tau \sim_s \tau'$ .  
1207 By (D.10),  $\forall s \in \mathcal{S}, \forall \tau, \tau' \in \{\pm 1\}^{\mathcal{S}}$  with  $\tau \sim_s \tau'$ ,

$$\begin{aligned}
 \text{SubOpt}_s(\widehat{\pi}; \tau) + \text{SubOpt}_s(\widehat{\pi}; \tau') &= \eta^{-1} \mathbf{KL}(\widehat{\pi} \parallel \pi_\tau^*) + \eta^{-1} \mathbf{KL}(\widehat{\pi} \parallel \pi_{\tau'}^*) \\
 &= 2\eta^{-1} \sum_{a \in \mathcal{A}} \widehat{\pi}(a|s) \log \frac{\widehat{\pi}(a|s)}{\sqrt{\pi_\tau^*(a|s) \pi_{\tau'}^*(a|s)}} \\
 &= 2\eta^{-1} \mathbf{KL}(\widehat{\pi}(\cdot|s) \parallel \bar{\pi}_{\tau, \tau'}(\cdot|s)) - 2\eta^{-1} \mathbb{E}_{a \sim \widehat{\pi}(\cdot|s)} \log \left( \sum_{b \in \mathcal{A}} \sqrt{\pi_\tau^*(b|s) \pi_{\tau'}^*(b|s)} \right) \\
 &\geq -2\eta^{-1} \log \left( \sum_{b \in \mathcal{A}} \sqrt{\pi_\tau^*(b|s) \pi_{\tau'}^*(b|s)} \right) \\
 &= \frac{1}{\eta} \log \frac{(\exp(\eta\delta) + 1)(\exp(-\eta\delta) + 1)}{4}, \tag{D.11}
 \end{aligned}$$

1222 where  $\bar{\pi}(\cdot|s) = \sqrt{\pi_\tau^*(\cdot|s) \pi_{\tau'}^*(\cdot|s)} / \sum_{b \in \mathcal{A}} \sqrt{\pi_\tau^*(b|s) \pi_{\tau'}^*(b|s)}$  for every  $s \in \mathcal{S}$ , the inequality is due  
1223 to the non-negativity of KL divergence, and the last equality follows from (D.7) together with the  
1224 design choice  $C - 1 = \exp(\eta\alpha)$ .

1226 **Case  $\eta\delta \leq 2$ .** Recall that  $\forall x \in \mathbb{R}, (e^x + e^{-x})/2 - 1 = x^2 \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k+2)!} \geq x^2/2$ , which implies  
1227

$$\text{(D.11)} = \frac{1}{\eta} \log \left( 1 + \frac{1}{2} \left( \frac{e^{\eta\delta} + e^{-\eta\delta}}{2} - 1 \right) \right) \geq \frac{1}{\eta} \log \left( 1 + \frac{\eta^2\delta^2}{4} \right) \geq \frac{1}{\eta} \cdot \frac{\eta^2\delta^2/4}{2} = \eta\delta^2/8. \tag{D.12}$$

1232 Here, the last inequality is due to  $\eta^2\delta^2/4 \leq 1$  and  $\forall x \in [0, 1], \log(1+x) \geq x/2$ .

1234 **Case  $\eta\delta > 2$ .** We have  $-\eta^{-1}2\log 2 \geq -\delta\log 2$ , which implies the following bound.

$$\text{(D.11)} \geq \frac{1}{\eta} \log \frac{\exp(\eta\delta) + 1}{4} \geq \frac{\eta\delta - 2\log 2}{\eta} = \delta - \eta^{-1}2\log 2 \geq (1 - \log 2)\delta \geq 3\delta/10. \tag{D.13}$$

1238 In summary, (D.12) and (D.13) imply that  $\forall s \in \mathcal{S}, \forall \tau, \tau' \in \{\pm 1\}^{\mathcal{S}}$  with  $\tau \sim_s \tau'$ ,

$$\text{SubOpt}_s(\widehat{\pi}; \tau) + \text{SubOpt}_s(\widehat{\pi}; \tau') \geq \frac{\eta\delta^2}{8} \wedge \frac{3\delta}{10}. \tag{D.14}$$

Let  $P_\tau$  be the distribution of  $(s, a, y)$  where  $s \sim \rho, a \sim \pi^{\text{ref}}(\cdot|s)$ , and  $y \sim \text{Bern}(r_\tau(s, a))$ . Then  $\forall x \in \mathcal{S} \forall \tau, \tau' \in \{\pm 1\}^{\mathcal{S}}$  with  $\tau \sim_x \tau'$ ,

$$\begin{aligned} \text{KL}(P_\tau \| P_{\tau'}) &= \frac{1}{S} \sum_{s,a} \pi^{\text{ref}}(a|s) \text{KL}(\text{Bern}(r_\tau(s, a)) \| \text{Bern}(r_{\tau'}(s, a))) \\ &= \frac{1}{S} \cdot C^{-1} \text{KL}(\text{Bern}(r_\tau(x, -1)) \| \text{Bern}(r_{\tau'}(x, -1))) \\ &\leq \frac{4\delta^2}{SC(0.25 - \delta^2)} \leq \frac{16\delta^2}{3SC}, \end{aligned} \quad (\text{D.15})$$

where we use the requirement  $\delta \leq 1/4$  and  $\text{KL}(\text{Bern}(p) \| \text{Bern}(q)) \leq (p - q)^2 / (q(1 - q))$ . Then let  $P_{\mathcal{D}_\tau}$  be the distribution of  $\mathcal{D}$  given the mean reward function  $r_\tau$ , we employ (D.15) to get

$$\text{KL}(P_{\mathcal{D}_\tau} \| P_{\mathcal{D}_{\tau'}}) = n \text{KL}(P_\tau \| P_{\tau'}) \leq \frac{16n\delta^2}{3SC}. \quad (\text{D.16})$$

Since  $n \geq 16SC^* = 16SC$  by design, we can set  $\delta = \sqrt{SC/n}$  (which ensures  $\delta \leq 1/4$ ) to obtain

$$\begin{aligned} \sup_{\text{inst}} \text{SubOpt}(\hat{\pi}; \text{inst}) &\geq \sup_{\tau \in \{\pm 1\}^{\mathcal{S}}} \text{SubOpt}(\hat{\pi}; \tau) \\ &\geq \frac{1}{S} \cdot S \cdot \frac{1}{4} \cdot \left( \frac{\eta\delta^2}{8} \wedge \frac{3\delta}{10} \right) \min_{\tau \sim \tau'} \exp \left( -\text{KL}(P_{\mathcal{D}_\tau} \| P_{\mathcal{D}_{\tau'}}) \right) \\ &\geq \left( \frac{\eta SC^*}{32n} \wedge \frac{3\sqrt{SC^*}}{40\sqrt{n}} \right) \exp(-16/3) \gtrsim \frac{\eta SC^*}{n} \wedge \sqrt{\frac{SC^*}{n}}. \end{aligned}$$

where the  $S^{-1}$  in the second inequality comes from (D.9), the second inequality is by substituting (D.14) into Assouad's Lemma (Lemma H.3), and the last inequality is due to (D.16).  $\square$

## E MISSING PROOF FROM SECTION 3

### E.1 PROOF OF THEOREM 3.2

Before coming to the proof, we first introduce some useful properties. The following properties characterize the convexity of  $f$ -divergence when  $f$  is (strongly) convex.

The strong-convexity of  $f$  implies that the corresponding  $f$ -divergence,  $D_f(\cdot \| \pi^{\text{ref}})$  is also strongly convex with respect to all  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$  supported by  $\pi^{\text{ref}}$ .

**Proposition E.1.** Given context  $s$ ,  $D_f(\pi(\cdot|s) \| \pi^{\text{ref}}(\cdot|s))$  is strict convex with respect to  $\pi$  if  $f$  is strictly convex.

**Proposition E.2.** Given context  $s$ ,  $\pi(\cdot|s) \mapsto D_f(\pi(\cdot|s) \| \pi^{\text{ref}}(\cdot|s))$  is  $4\alpha$ -strong convex with respect to the metric TV if  $f$  is  $\alpha$ -strongly convex.

*Proof of Proposition E.2.* We first show the gradient of  $D_f$  with respect to  $\pi$ .

$$\frac{\partial D_f(\pi \| \pi^{\text{ref}})}{\pi(a)} = \frac{\partial}{\partial \pi(a)} \sum_{b \in \mathcal{A}} \pi^{\text{ref}}(b) f \left( \frac{\pi(b)}{\pi^{\text{ref}}(b)} \right) = f' \left( \frac{\pi(a)}{\pi^{\text{ref}}(a)} \right).$$

1296 Now consider  $\pi_1, \pi_2 \in \Delta(\mathcal{A})$  supported by  $\pi^{\text{ref}}$ .  
1297

$$\begin{aligned}
1298 \quad & D_f(\pi_1 || \pi^{\text{ref}}) - D_f(\pi_2 || \pi^{\text{ref}}) - \langle \pi_1 - \pi_2, \nabla D_f(\pi_2 || \pi^{\text{ref}}) \rangle \\
1299 \quad &= \sum_{a \in \mathcal{A}} \pi^{\text{ref}}(a) \left( f\left(\frac{\pi_1(a)}{\pi^{\text{ref}}(a)}\right) - f\left(\frac{\pi_2(a)}{\pi^{\text{ref}}(a)}\right) \right) - \sum_{a \in \mathcal{A}} (\pi_1(a) - \pi_2(a)) f'\left(\frac{\pi_2(a)}{\pi^{\text{ref}}(a)}\right) \\
1300 \quad &= \sum_{a \in \mathcal{A}} \pi^{\text{ref}}(a) \left( f\left(\frac{\pi_1(a)}{\pi^{\text{ref}}(a)}\right) - f\left(\frac{\pi_2(a)}{\pi^{\text{ref}}(a)}\right) - \left(\frac{\pi_1(a)}{\pi^{\text{ref}}(a)} - \frac{\pi_2(a)}{\pi^{\text{ref}}(a)}\right) f'\left(\frac{\pi_2(a)}{\pi^{\text{ref}}(a)}\right) \right) \\
1301 \quad &\geq \frac{\alpha}{2} \sum_{a \in \mathcal{A}} \pi^{\text{ref}}(a) \left( \frac{\pi_1(a)}{\pi^{\text{ref}}(a)} - \frac{\pi_2(a)}{\pi^{\text{ref}}(a)} \right)^2 \\
1302 \quad &= \frac{\alpha}{2} \sum_{a \in \mathcal{A}} \frac{1}{\pi^{\text{ref}}(a)} (\pi_1(a) - \pi_2(a))^2 \\
1303 \quad &\geq \frac{\alpha}{2} \left( \sum_{a \in \mathcal{A}} |\pi_1(a) - \pi_2(a)| \right)^2,
\end{aligned}$$

1313 where the first inequality holds due to  $f$ 's strong convexity and the second holds due to  
1314 Cauchy–Schwarz. The proof finishes since  $\|\pi_1 - \pi_2\|_1 = 2\text{TV}(\pi_1 || \pi_2)$ .  $\square$   
1315

1316 We first introduce some notation and important properties concerning the convex conjugate  
1317 of functions. Given some context  $s$ , we denote the regularization term as  $H_s(\pi) =$   
1318  $\eta^{-1} D_f(\pi(\cdot|s) || \pi^{\text{ref}}(\cdot|s))$ . We use  $H_s^*(r)$  to denote the convex conjugate of  $H_s$ , which is defined  
1319 as

$$1320 \quad H_s^*(r) = \sup_{\pi \in \mathcal{S} \rightarrow \Delta^{|\mathcal{A}|}} \{ \langle \pi(\cdot|s), r(s, \cdot) \rangle - H_s(\pi) \}.$$

1323 We have the following properties for the convex conjugate. The first property gives the gradient of  
1324 convex conjugate (see, e.g., Zhou 2018, Lemma 5).

1325 **Proposition E.3.** Given context  $s$ , and convex  $f$ , let  $\pi_r \in \text{argmax}_{\pi} \{ \langle \pi(\cdot|s), r(s, \cdot) \rangle - H_s(\pi) \}$  for  
1326 some  $r$ , then the gradient of  $H_s^*$  is given by  $\nabla H_s^*(r) = \pi_r(\cdot|s)$ .

1327 We also need some properties of  $\nabla^2 H_s^*$ , the Hessian matrix of the convex conjugate function. We  
1328 first give the Hessian matrix of the original function  $H_s$  as follows.

$$1330 \quad \nabla^2 H_s(\pi) = \eta^{-1} \text{diag} \left( \frac{f''\left(\frac{\pi(a_1|s)}{\pi^{\text{ref}}(a_1|s)}\right)}{\pi^{\text{ref}}(a_1|s)}, \dots, \frac{f''\left(\frac{\pi(a_{|\mathcal{A}|}|s)}{\pi^{\text{ref}}(a_{|\mathcal{A}|}|s)}\right)}{\pi^{\text{ref}}(a_{|\mathcal{A}|}|s)} \right). \quad (\text{E.1})$$

1333 Furthermore, when  $f$  is  $\alpha$ -strongly convex, we have

$$1335 \quad \nabla^2 H_s(\pi) \succeq \alpha \eta^{-1} \text{diag}(\pi^{\text{ref}}(a_1|s)^{-1}, \dots, \pi^{\text{ref}}(a_{|\mathcal{A}|}|s)^{-1}).$$

1337 The following lemma, which gives an estimate of  $\nabla^2 H_s^*$ , is the pivot of the proof.

1338 **Lemma E.4.** For any reward  $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ , we have

$$1340 \quad \nabla^2 H_s^*(r) \preceq \alpha^{-1} \eta \text{diag}(\pi^{\text{ref}}(a_1|s), \dots, \pi^{\text{ref}}(a_{|\mathcal{A}|}|s)).$$

1342 *Proof of Lemma E.4.* Given reward function  $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ , we consider

$$1344 \quad \pi_r \in \text{argmax}_{\pi \in \mathcal{S} \rightarrow \Delta^{|\mathcal{A}|}} \{ \langle \pi(\cdot|s), r(\cdot|s) \rangle - H_s(\pi) \}.$$

1346 From (E.1) we know that  $\nabla^2 H_s(\pi_r)$  is invertible. Therefore, by Penot 1994, Proposition 3.2, we  
1347 have  $\nabla^2 H_s^*(r) \preceq (\nabla^2 H_s(\pi_r))^{-1}$ . Since  $f$  is  $\alpha$ -strongly convex, we have

$$1348 \quad \nabla^2 H_s^*(r) \preceq \alpha^{-1} \eta \text{diag}(\pi^{\text{ref}}(a_1|s), \dots, \pi^{\text{ref}}(a_{|\mathcal{A}|}|s)),$$

1349 which finishes the proof.  $\square$

1350 Now we are ready to prove Theorem 3.2.  
 1351

1352 *Proof of Theorem 3.2.* Consider our estimation  $\bar{g}$  which approximates the ground truth reward func-  
 1353 tion  $g^*$ , we know that  
 1354

$$1355 \hat{\pi} = \operatorname{argmax}_{\pi \in \mathcal{S} \rightarrow \Delta(\mathcal{A})} \left\{ \mathbb{E}_{(s,a) \sim \rho \times \pi} [\bar{g}(s,a)] - \eta^{-1} \mathbb{E}_{s \sim \rho} [D_f(\pi \parallel \pi^{\text{ref}})] \right\}. \\ 1356$$

1357 We have the following sub-optimality decomposition  
 1358

$$1359 J(\pi^*) - J(\hat{\pi}) = \mathbb{E}_{s \sim \rho} \left[ \mathbb{E}_{a \sim \pi^*} [g^*(s,a)] - \mathbb{E}_{a \sim \hat{\pi}} [g^*(s,a)] - \eta^{-1} [D_f(\pi^* \parallel \pi^{\text{ref}}) - D_f(\hat{\pi} \parallel \pi^{\text{ref}})] \right] \\ 1360 = \mathbb{E}_{s \sim \rho} [H_s^*(g^*) - H_s^*(\bar{g}) - \langle \hat{\pi}, g^* - \bar{g} \rangle] \\ 1361 = \mathbb{E}_{s \sim \rho} [H_s^*(g^*) - H_s^*(\bar{g}) - \langle \nabla H_s^*(\bar{g}), g^* - \bar{g} \rangle] \\ 1362 = \mathbb{E}_{s \sim \rho} [(g^* - \bar{g})^\top \nabla^2 H_s^*(\bar{g})(g^* - \bar{g})], \\ 1363$$

1364 where  $\tilde{g} = \gamma g^* + (1 - \gamma)\bar{g}$  and  $\gamma \in [0, 1]$  and the last equation holds due to Taylor's expansion.  
 1365 Now, for any  $\delta \in (0, 1)$  and  $\epsilon_c > 0$ , with probability at least  $1 - \delta$   
 1366

$$1367 J(\pi^*) - J(\hat{\pi}) = \mathbb{E}_{s \sim \rho} [(g^* - \bar{g})^\top \nabla^2 H_s^*(\bar{g})(g^* - \bar{g})] \\ 1368 \leq \alpha^{-1} \eta \mathbb{E}_{s \sim \rho} [(g^* - \bar{g})^\top \operatorname{diag}(\pi^{\text{ref}}(a_1|s), \dots, \pi^{\text{ref}}(a_{|\mathcal{A}|}|s))(g^* - \bar{g})] \\ 1369 = \alpha^{-1} \eta \mathbb{E}_{(s,a) \sim \rho \times \pi^{\text{ref}}} [(g^*(s,a) - \bar{g}(s,a))^2] \\ 1370 \leq \alpha^{-1} \eta \left( \frac{128}{3n} \log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 18\epsilon_c \right), \\ 1371$$

1372 where the first inequality holds due to Lemma E.4 and last inequality holds due to equation (D.2).  
 1373 Setting  $\epsilon_c = O(n^{-1})$  completes the proof.  $\square$   
 1374

## 1377 E.2 PROOF OF THEOREM 3.4

1378 We first provide the following lemma that gives the close form of optimal policy under  $\chi^2$ -divergence  
 1379 regularization.  
 1380

1381 **Lemma E.5** (Huang et al. (2025a, Lemma G.2)). Let  $\pi^*$  be the optimal policy of  $\chi^2$ -divergence  
 1382 regularized objective with reward function  $r$ , then  $\pi^*$  has the closed form  
 1383

$$1384 \pi^*(\cdot) = \pi^{\text{ref}}(\cdot) \max \{0, \eta(r(\cdot) - \lambda)\}, \text{ where } \sum_{a \in \mathcal{A}} \pi_{f\text{div}}^*(a) = 1. \\ 1385$$

1386 By Proposition E.2,  $\pi_{f\text{div}}^* = \operatorname{argmax}_{\pi \in \Delta(\mathcal{A})} J_{f\text{div}}(\pi)$  is unique. The sub-optimality gap for  $f$ -  
 1387 divergence is consequently defined as  
 1388

$$1389 \text{SubOpt}_{f\text{div}}(\cdot) := \text{SubOpt}_{f\text{div}}(\cdot; \mathcal{A}, r, \pi^{\text{ref}}) = J_{f\text{div}}(\pi_{f\text{div}}^*) - J_{f\text{div}}(\cdot). \quad (\text{E.2}) \\ 1390$$

1391 Now we are ready to prove Theorem 3.4.  
 1392

1393 *Proof of Theorem 3.4.* We still consider the family of contextual bandits  $\text{CB}_{\mathcal{G}}$  given by (D.5). We,  
 1394 still, aim to prove the following statement. Fixing any  $S \geq 32 \log 2$ ,  $\eta > 4 \log 2$  and  $\alpha$ , we set  
 1395  $f(x) := \alpha(x - 1)^2/2$ , then for any estimator  $\mathcal{D} \mapsto \hat{\pi} \in \Delta(\mathcal{A}|\mathcal{S})$ , for any  $n$  sufficiently large,  
 1396 there exist some function class  $\mathcal{G}$ , such that  $\exists \text{inst} = (\mathcal{S}, \mathcal{A}, \rho, r, \pi^{\text{ref}}, \eta) \in \text{CB}_{\mathcal{G}}$  with  $|\mathcal{S}| = S =$   
 1397  $\Theta(\log |\mathcal{G}|)$ , and  
 1398

$$\text{SubOpt}_{f\text{div}}(\hat{\pi}; \text{inst}) \gtrsim \alpha^{-1} \eta S n^{-1}. \quad (\text{E.3})$$

1399 Since  $\log |\mathcal{G}| \geq \log \mathcal{N}_{\mathcal{G}}(\epsilon)$  for any  $\epsilon \in (0, 1)$ , equation (E.3) yields the desired bound.  
 1400

1401 We again omit subscripts  $f\text{div}$  when it is clear in context. We set  $\mathcal{S} = [S]$ ,  $\mathcal{A} = \{-1, +1\}$ , and  
 1402  $\rho = \text{Unif}(\mathcal{S})$ . For all  $s \in \mathcal{S}$ ,  $\pi^{\text{ref}} = \text{Unif}(\mathcal{A})$ . We further consider the following reward function  
 1403 class. We leverage Lemma H.4 and obtain a set  $\mathcal{V} \in \{-1, +1\}^S$  such that (1)  $|\mathcal{V}| \geq \exp(S/8)$   
 1404 and (2) for any  $v, v' \in \mathcal{V}, v \neq v'$ , one has  $\|v - v'\|_1 \geq S/2$ . We construct the following reward  
 1405

function class where the reward follows Bernoulli distribution and the mean functions are given by the function class

$$\mathcal{G} = \{r_v(s, -1) = 1/2 + v_s \delta, r_{v'}(s, +1) = 1/2 + v'_s \delta, \forall s \in \mathcal{S} | v \in \mathcal{V}\},$$

where  $\delta \in (0, \eta^{-1} \alpha]$  is to be specified later. Fix some context  $s$  and  $v_1 \neq v_2$  different at entry  $s$  and corresponding reward  $r_1$  and  $r_2$ . Without loss of generality, we assume  $r_1(s, \cdot) = (1/2 + \delta, 1/2 - \delta)$  and  $r_2(s, \cdot) = (1/2 - \delta, 1/2 + \delta)$ . Then direct calculation implies that

$$\begin{aligned}\pi_1^*(\cdot | s) &= \frac{1}{2} \max\{0, \eta \alpha^{-1}(r_1(s, \cdot) - \lambda)\} = 0.5 \eta \alpha^{-1}(r_1(s, \cdot) - \lambda), \\ \pi_2^*(\cdot | s) &= \frac{1}{2} \max\{0, \eta \alpha^{-1}(r_2(s, \cdot) - \lambda)\} = 0.5 \eta \alpha^{-1}(r_2(s, \cdot) - \lambda),\end{aligned}$$

where  $\lambda = 0.5 - \eta^{-1} \alpha$ . Note that  $2\chi^2(\mu \| \nu) + 1 = \sum_{a \in \mathcal{A}} [\mu(a)]^2 / \nu(a)$  and  $\chi^2 = D_f$ , we obtain that  $\forall \widehat{\pi}$ ,

$$\text{SubOpt}_s(\widehat{\pi}(\cdot | s); r_1) + \text{SubOpt}_s(\widehat{\pi}(\cdot | s); r_2) \quad (\text{E.4})$$

$$\begin{aligned}&= \langle r_1(s, \cdot), \pi_1^*(\cdot | s) \rangle + \langle r_2(s, \cdot), \pi_2^*(\cdot | s) \rangle - \overbrace{\langle r_1(s, \cdot) + r_2(s, \cdot), \widehat{\pi}(\cdot | s) \rangle}^{=1} + \overbrace{2\eta^{-1} \alpha \chi^2(\widehat{\pi}(\cdot | s) \| \pi^{\text{ref}}(\cdot | s))}^{\geq 0} \\ &\quad - \eta^{-1} \alpha \cdot \chi^2(\pi_1^*(\cdot | s) \| \pi^{\text{ref}}(\cdot | s)) - \eta^{-1} \alpha \cdot \chi^2(\pi_2^*(\cdot | s) \| \pi^{\text{ref}}(\cdot | s)) \\ &\geq 2\langle r_1(s, \cdot), \pi_1^*(\cdot | s) \rangle - 1 - 2\eta^{-1} \alpha \cdot \chi^2(\pi_1^*(\cdot | s) \| \pi^{\text{ref}}(\cdot | s)) \\ &= 1 + \frac{2\eta\delta^2}{\alpha} - 1 - \frac{\eta\delta^2}{\alpha} = \frac{\eta\delta^2}{\alpha}. \quad (\text{E.5})\end{aligned}$$

Now we take expectation over all possible contexts and recall that  $\|v - v'\|_1 \geq S/2$  for  $v \neq v'$ , we know that for any  $r_1 \neq r_2 \in \mathcal{G}$

$$\text{SubOpt}(\widehat{\pi}; r_1) + \text{SubOpt}(\widehat{\pi}; r_2) \geq \frac{\eta\delta^2}{2\alpha}$$

Given any mean reward function  $r \in \mathcal{G}$ , let  $P_r$  be the distribution of  $(s, a, \mathbf{r})$  when  $s \sim \rho$ ,  $a \sim \pi^{\text{ref}}(\cdot | s)$ , and  $\mathbf{r} \sim \text{Bern}(r(s, a))$ . Suppose  $P_{\mathcal{D}_r}$  is the distribution of the dataset given mean reward function  $r$ , then  $\text{KL}(P_{\mathcal{D}_{r_1}} \| P_{\mathcal{D}_{r_2}}) = n\text{KL}(P_{r_1} \| P_{r_2})$  for any pair of  $r_1, r_2 \in \mathcal{G}$ . Now we invoke Fano's inequality (Lemma H.2) to obtain

$$\begin{aligned}\inf_{\pi} \sup_{\text{inst} \in \text{CB}_{\mathcal{G}}} \text{SubOpt}(\widehat{\pi}; \text{inst}) &\geq \frac{\eta\delta^2}{4\alpha} \left( 1 - \frac{\max_{r_1 \neq r_2 \in \mathcal{G}} \text{KL}(P_{\mathcal{D}_{r_1}} \| P_{\mathcal{D}_{r_2}}) + \log 2}{\log |\mathcal{G}|} \right) \\ &\geq \frac{\eta\delta^2}{4\alpha} \left( 1 - \frac{64n\delta^2 + 8\log 2}{S} \right),\end{aligned}$$

where the second inequality holds due to  $\text{KL}(\text{Bern}(p) \| \text{Bern}(q)) \leq (p - q)^2 / [q(1 - q)]$ . Let  $\delta = 16^{-1} \sqrt{S n^{-1}}$ , then we obtain that for all  $\pi$  we have

$$\sup_{\text{inst} \in \text{CB}_{\mathcal{G}}} \text{SubOpt}(\widehat{\pi}; \text{inst}) \gtrsim \frac{\eta S}{\alpha n},$$

which finishes the proof in that  $\log_2 |\mathcal{G}| = S$ .  $\square$

## F GENERALIZATION TO CONTEXTUAL DUELING BANDITS

In this section, we extend our algorithm to the problems of regularized contextual dueling bandits, where the learner receives preference comparison instead of absolute signals. Our setup largely follows Zhu et al. (2023); Zhan et al. (2023) and the notion of sub-optimality follows Xiong et al. (2024); Zhao et al. (2024).

---

1458 **Algorithm 3** Offline KL-Regularized Pessimistic Contextual Dueling Bandit (KL-PCDB)  
1459  
1460 **Require:** regularization  $\eta$ , reference policy  $\pi^{\text{ref}}$ , function class  $\mathcal{G}$ , offline dataset  $\mathcal{D} =$   
1461  $\{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$   
1462 1: Compute the maximum likelihood estimator of the reward function  
1463  $\bar{g} = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^n \left[ y_i \log \sigma \left( [g(s_i, a_i^1) - g(s_i, a_i^2)] \right) + (1 - y_i) \log \sigma \left( [g(s_i, a_i^2) - g(s_i, a_i^1)] \right) \right]$   
1464  
1465 2: Let  $\hat{g}(s, a) = \bar{g}(s, a) - \Gamma_n(s, a)$ , where  $\Gamma_n(s, a)$  is the bonus term in (F.1)  
1466  
1467 **Ensure:**  $\hat{\pi}(a|s) \propto \pi^{\text{ref}}(a|s) \exp(\eta \cdot \hat{g}(s, a))$

---

1469  
1470 F.1 PROBLEM SETUP  
1471

1472 We still consider contextual bandits  $(\mathcal{S}, \mathcal{A}, r, \pi^{\text{ref}})$  where  $\mathcal{S}$  is the state space,  $\mathcal{A}$  is the action space  
1473 and  $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$  is the reward function.<sup>6</sup> But only relative preference feedback is available,  
1474 viz., we have an i.i.d. offline dataset  $\mathcal{D} = \{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$ , where  $s_i \in \mathcal{S}$  is generated from  
1475 distribution  $\rho$  and  $a_i^1, a_i^2 \sim \pi^{\text{ref}}$ . The binary preference label  $y_i = 1$  indicates  $a_i^1$  is preferred over  $a_i^2$   
1476 (denoted by  $a^1 \succ a^2$ ) and 0 for  $a^2 \succ a^1$  given context  $s$ . In this work we consider the Bradley-Terry  
1477 Model, where  $\mathbb{P}[y = 1|s, a^1, a^2] = \sigma(r(s_i, a_i^1) - r(s_i, a_i^2))$ , where  $\sigma(x) = (1 + e^{-x})^{-1}$  is the  
1478 link function. The objective here identical to (2.1) for KL-regularization and (3.1) for  $f$ -divergence  
1479 regularization. Our goal is still to find an  $\epsilon$ -optimal policy. To control the complexity of the function  
1480 class  $\mathcal{G}$ , we assume that Assumption 2.1 still holds here.

1481 **Concentrability.** Analogous to Section 2, we need our estimation from offline dataset generalizable  
1482 to the state-action pairs visited by our obtained policy. While density-ratio-based concentrability  
1483 can be directly adapted to dueling bandit, we need a slightly different notion of  $D^2$ -divergence.  
1484 This is because in dueling bandit, we cannot observe the absolute reward and best estimation  $g$   
1485 we can achieve is that for any state  $s$  and actions  $a^1, a^2$ , our estimated  $g(s, a^1) - g(s, a^2) \approx$   
1486  $r(s, a^1) - r(s, a^2)$ . This implies that there exists some mapping  $b : \mathcal{S} \rightarrow [-1, 1]$  such that  
1487  $g(s, a) - b(s) \approx r(s, a)$  on the offline data, which leads to the following definition.

1488 **Definition F.1.** Given a class of functions  $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R})$  and some policy  $\pi$ , let  $\mathcal{B} = (\mathcal{S} \rightarrow$   
1489  $[-1, 1])$  be the function class, define the  $D^2$ -divergence  $D_{\mathcal{G}}^2((s, a); \pi)$  as

$$1491 \sup_{g, h \in \mathcal{G}} \inf_{b \in \mathcal{B}} \frac{(g(s, a) - h(s, a) - b(s))^2}{\mathbb{E}_{s \sim \rho} \operatorname{Var}_{a' \sim \pi(\cdot|s')} [g(s', a') - h(s', a')]}.$$

1494 A similar definition has been introduced in Zhao et al. (2024, Definition 2.6), which underpins the  
1495 following two assumptions that characterize the coverage ability of  $\pi^{\text{ref}}$  similarly as in Section 2.

1496 Given a reference policy  $\pi^{\text{ref}}$ , we define two coverage notions for contextual dueling bandits.

1497 **Assumption F.2** (All-policy concentrability).  $D^2 := \sup_{(s, a) \in \mathcal{S} \times \mathcal{A}} D_{\mathcal{G}}^2((s, a); \pi^{\text{ref}}) < \infty$ .

1499 **Assumption F.3** (Single-policy concentrability).  $D_{\pi^*}^2 := \mathbb{E}_{(s, a) \sim \rho \times \pi^*} [D_{\mathcal{G}}^2((s, a); \pi^{\text{ref}})] < \infty$ .

1500 Similar single-policy concentrability assumptions have appeared in previous work in offline context-  
1501 ual dueling bandits (Huang et al., 2025b; Song et al., 2024) and similar notions has also appeared  
1502 in the analysis of model-based RL (Uehara & Sun, 2021; Wang et al., 2024). Still, while Assump-  
1503 tion F.3 is strictly weaker than Assumption F.2, in general cases, the two quantities,  $C^{\pi^*}$  and  $D_{\pi^*}^2$   
1504 cannot be bounded by each other.

1505  
1506 F.2 ALGORITHMS AND RESULTS  
1507

1508 F.2.1 ALGORITHMS FOR KL-REGULARIZED CONTEXTUAL DUELING BANDITS

1509 We elucidate KL-PCDB for offline KL-regularized contextual dueling bandits, whose pseudocode is  
1510 summarized in Algorithm 3. KL-PCDB first estimate the ground truth function  $g^*$  on offline dataset

1511 <sup>6</sup>We overload some notations in Section 2 by their dueling counterparts for notational simplicity.

---

**Algorithm 4** Offline  $f$ -Divergence Regularized Contextual Dueling Bandits ( $f$ -CDB)

---

**Require:** regularization  $\eta$ , reference policy  $\pi^{\text{ref}}$ , function class  $\mathcal{G}$ , offline dataset  $\mathcal{D} = \{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$

1: Compute the maximum likelihood estimator of the reward function

$$\bar{g} = \operatorname{argmin}_{g \in \mathcal{G}} \sum_{i=1}^n \left[ y_i \log \sigma\left(\left[g(s_i, a_i^1) - g(s_i, a_i^2)\right]\right) + (1 - y_i) \log \sigma\left(\left[g(s_i, a_i^2) - g(s_i, a_i^1)\right]\right) \right].$$

2: Compute the optimal policy with respect to reward  $\bar{g}$

$$\hat{\pi}(\cdot|s) \leftarrow \operatorname{argmax}_{\pi(\cdot|s) \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} \pi(a|s) \bar{g}(s, a) + \eta^{-1} D_f(\pi(\cdot|s) \| \pi^{\text{ref}}(\cdot|s))$$

**Ensure:**  $\hat{\pi}(a|s)$

with maximum likelihood estimator (MLE) to estimate a function  $\bar{g} \in \mathcal{G}$ . After that, analogous to Algorithm 1, we adopt the principle of pessimism in the face of uncertainty. Specifically, we define the penalty term

$$\Gamma_n(s, a) = \beta \sqrt{D_{\mathcal{G}}^2((s, a), \pi^{\text{ref}})}, \quad (\text{F.1})$$

where

$$\beta^2 = 128 \log(2\mathcal{N}_G(\epsilon_c)/\delta)/3n + 18\epsilon_c = \tilde{O}(n^{-1}) \quad (\text{F.2})$$

and then subtract it from the MLE  $\bar{g}$  to obtain a pessimistic estimator  $\hat{g}$ . KL-PCB then output the policy  $\hat{\pi}$ , maximizing the estimated objective

$$\widehat{J}(\pi) = \mathbb{E}_{(s,a) \sim \rho \times \pi} \left[ \widehat{g}(s,a) - \eta^{-1} \log \frac{\pi(a|s)}{\pi^{\text{ref}}(a|s)} \right],$$

the maximizer of which is in closed form as the counterpart of (2.2).

$$\widehat{\pi}(a|s) \propto \pi^{\text{ref}}(a|s) \exp\left(\eta \cdot \widehat{q}(s, a)\right),$$

We provide the following theoretical guarantees for Algorithm 3.

**Theorem F.4.** Under Assumption F.3, if we set  $\Gamma_n$  according to (F.1), then for sufficiently small  $\epsilon \in (0, 1)$ , with probability at least  $1 - \delta$ ,  $n = \tilde{O}(\eta(D_{\pi^*}^2 \wedge C^{\pi^*})\epsilon^{-1})$  is sufficient to guarantee the output policy  $\hat{\pi}$  of Algorithm 3 to be  $\epsilon$ -optimal.

**Remark F.5.** Zhao et al. (2024) achieved an  $\tilde{O}(\epsilon^{-1})$  sample complexity under Assumption F.2. Comparing to Zhao et al. (2024), KL-PCDB achieves the same  $\tilde{O}(\epsilon^{-1})$  sample complexity but only requiring Assumption F.3, which is weaker than Assumption F.2.

The following theorem provides the sample complexity lower bound for KL-regularized dueling contextual bandits.

**Theorem F.6.** For any sufficiently small  $\epsilon \in (0, 1)$ ,  $\eta > 0$ ,  $1 \leq C^* \leq \exp(\eta/2)/2$ , and any algorithm  $\text{Alg}$ , there is a KL-regularized contextual dueling bandit instance with single-policy concentrability  $C^{\pi^*} \leq C^*$  such that  $\text{Alg}$  requires at least  $\Omega\left(\min\{\eta C^* \log \mathcal{N}_{\mathcal{G}}(\epsilon_c)/\epsilon, \log \mathcal{N}_{\mathcal{G}}(\epsilon_c)(C^*)^2/\epsilon^2\}\right)$  samples to return an  $\epsilon$ -optimal policy.

**Remark F.7.** Theorem F.6 shows that when  $\epsilon$  is sufficiently small, any algorithm for offline KL-regularized contextual dueling bandits requires at least  $\Omega(\eta C^{\pi^*} \log \mathcal{N}_{\mathcal{G}}(\epsilon) \epsilon^{-1})$  samples to output an  $\epsilon$ -optimal policy, which matches the sample complexity upper bound in Theorem F.4, indicating that KL-PCB is nearly optimal.

### F.2.2 ALGORITHM AND RESULTS FOR $f$ -DIVERGENCE REGULARIZED CDBS

We present an offline learning algorithm for  $f$ -divergence regularized contextual dueling bandit,  $f$ -CDB, in Algorithm 4.  $f$ -CDB first leverages maximum likelihood estimator to find a function

1566  $\bar{g} \in \mathcal{G}$  that minimizes its risk on the offline dataset. Then the algorithm constructs the output policy  
 1567  $\hat{\pi}$  that maximizes the  $f$ -divergence regularized objective induced by  $\bar{g}$ . Similar to Algorithm 2,  
 1568 we do not require any pessimism in  $f$ -CDB. The following theorem provides an upper bound of  
 1569 Algorithm 4.

1570 **Theorem F.8.** For any sufficiently small  $\epsilon \in (0, 1)$ , and  $\eta, \alpha > 0$ , with probability at least  $1 - \delta$ ,  
 1571  $n = \tilde{O}(\alpha^{-1} \eta \log \mathcal{N}(\epsilon) \epsilon^{-1})$  is sufficient to guarantee that the output policy  $\hat{\pi}$  of Algorithm 4 is  
 1572  $\epsilon$ -optimal.

1573  
 1574 The following theorem provides a lower bound for offline  $f$ -divergence regularized contextual du-  
 1575 eling bandit with strongly convex  $f$ .

1576 **Theorem F.9.** For any  $\epsilon \in (0, 1)$ ,  $\alpha, \eta > 0$ , and offline RL algorithm  $\text{Alg}$ , there is an  $\alpha$ -strongly  
 1577 convex  $f$  and  $f$ -divergence regularized contextual dueling bandit instance such that  $\text{Alg}$  requires at  
 1578 least  $\Omega(\alpha^{-1} \eta \log \mathcal{N}(\epsilon) \epsilon^{-1})$  samples to return an  $\epsilon$ -optimal policy.

1579 **Remark F.10.** Theorem F.9 indicates that, when  $\epsilon$  is sufficiently small, to produce an  $\epsilon$ -optimal  
 1580 policy, any algorithm for offline  $f$ -regularized contextual bandits with strongly convex  $f$  requires  
 1581 at least  $\tilde{\Omega}(\alpha^{-1} \eta \epsilon^{-1})$  samples. This lower bound matches the sample-complexity upper bound in  
 1582 Theorem F.8, indicating that Algorithm 4 is nearly optimal.

1583

## G MISSING PROOF FROM APPENDIX F

1584

### G.1 PROOF OF THEOREM F.4

1585

1586 The proof follows the proof in Section 2. At the beginning, we first define the event  $\mathcal{E}(\delta)$  given  
 1587  $\delta > 0$  as

1588

$$\mathcal{E}(\delta) := \left\{ \exists b : \mathcal{S} \rightarrow [-1, 1], \forall (s, a) \in \mathcal{S} \times \mathcal{A}, |\bar{g}(s, a) - b(s) - g^*(s, a)| \leq \Gamma_n(s, a) \right\}. \quad (\text{G.1})$$

1589

1590 Here,  $\Gamma_n$  is defined in (F.1). We abuse the notation and define  $b(\cdot)$  as

1591

1592

$$b = \operatorname{argmin}_{\mathcal{B}} \sup_{(s, a) \in \mathcal{S} \times \mathcal{A}} \Phi_b(s, a) - \Gamma_n(s, a), \quad (\text{G.2})$$

1593

1594 where  $\Phi_b(s, a) = |\bar{g}(s, a) - b(s) - g^*(s, a)|$  and when  $\mathcal{E}$  holds, for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , we have  
 1595  $\Phi_b(s, a) \leq \Gamma_n(s, a)$ . This indicates that the least square estimation  $\bar{g}$  obtained in Line 1 of Algo-  
 1596 rithm 3, after adjusted by some bias function  $b$ , is close to the true function  $g^*$ . The following lemma  
 1597 shows that this event holds with high probability.

1598

1599 **Lemma G.1.** For any  $\delta > 0$ ,  $\mathbb{P}(\mathcal{E}(\delta)) \geq 1 - \delta$ .

1600

1601

1602 *Proof.* From Lemma H.1, we have that with probability at least  $1 - \delta$ , it holds that

1603

1604

$$\mathbb{E}_{s' \sim \rho} \operatorname{Var}_{a' \sim \pi^{\text{ref}}(\cdot | s')} [\bar{g}(s', a') - g^*(s', a')] \leq O\left(\frac{1}{n} \log(\mathcal{N}_G(\epsilon_c)/\delta) + \epsilon_c\right). \quad (\text{G.3})$$

1605

1606 It further holds true that for some  $b : \mathcal{S} \rightarrow \mathbb{R}$

1607

1608

1609

$$D_G^2((s, a), \pi^{\text{ref}}) \cdot \mathbb{E}_{s \sim \rho} \operatorname{Var}_{a \sim \pi^{\text{ref}}(\cdot | s)} [\bar{g}(s, a) - g^*(s, a)] \geq (\bar{g}(s, a) - b(s) - g^*(s, a))^2. \quad (\text{G.4})$$

Substituting (G.3) into (G.4), we have

1610

1611

$$\inf_b (\bar{g}(s, a) - b(s) - g^*(s, a))^2 \quad (\text{G.5})$$

1612

1613

1614

$$= \inf_b \frac{(\bar{g}(s, a) - b(s) - g^*(s, a))^2}{\mathbb{E}_{s' \sim \rho} \operatorname{Var}_{a' \sim \pi^{\text{ref}}(\cdot | s')} [\bar{g}(s', a') - g^*(s', a')]} \mathbb{E}_{s' \sim \rho} \operatorname{Var}_{a' \sim \pi^{\text{ref}}(\cdot | s')} [\bar{g}(s', a') - g^*(s', a')] \quad (\text{G.6})$$

1615

1616

1617

$$\leq D_G^2((s, a), \pi^{\text{ref}}) \mathbb{E}_{\pi^{\text{ref}}} [(\bar{g}(s, a) - b(s) - g^*(s, a))^2] \quad (\text{G.6})$$

1618

1619

$$\leq D_G^2((s, a), \pi^{\text{ref}}) O\left(\frac{1}{n} \log(\mathcal{N}_G(\epsilon_c)/\delta) + \epsilon_c\right), \quad (\text{G.7})$$

where the first inequality holds due to the definition of  $D_G^2((s, a), \pi^{\text{ref}})$  and the last inequality holds  
 due to Lemma H.1.  $\square$

1620 We overload the following quantities. For any  $\gamma \in [0, 1]$  and  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , we define  
 1621

$$1622 g_\gamma(s, a) := \gamma(\hat{g}(s, a) - b(s)) + (1 - \gamma)g^*(s, a).$$

1623 Furthermore, we introduce the following quantities  
 1624

$$1625 \pi_\gamma(\cdot | \cdot) = \pi_{g_\gamma}(\cdot | \cdot) \propto \pi^{\text{ref}}(\cdot | \cdot) \exp(\eta g_\gamma(\cdot, \cdot)),$$

$$1626 G(\gamma) := \mathbb{E}_{\rho \times \pi_\gamma} [(\hat{g}(s, a) - b(s) - g^*(s, a))^2],$$

1628 where  $b(\cdot)$  is defined in (G.2). We still have the monotonicity of the function  $G(\gamma)$ , which is char-  
 1629 acterized by the following lemma.

1630 **Lemma G.2.** On event  $\mathcal{E}(\delta)$ ,  $0 \in \text{argmax}_{\gamma \in [0, 1]} G(\gamma)$ .  
 1631

1632 *Proof.* For simplicity, we use  $\Delta(s, a)$  to denote  $\hat{g}(s, a) - b(s) - g^*(s, a)$  in *this* proof. Then on  
 1633 event  $\mathcal{E}(\delta)$ , we know that  $\Delta(s, a) \leq 0$  for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . Taking derivatives of  $G$  w.r.t.,  $\gamma$   
 1634 directly, we conclude that for all  $\gamma \in [0, 1]$ ,  
 1635

$$1636 G'(\gamma) = \eta \mathbb{E}_\rho \mathbb{E}_{a \sim \pi_\gamma} [\Delta^2(s, a) (\Delta(s, a) - \mathbb{E}_{a' \sim \pi_\gamma} [\Delta(s, a')])]$$

$$1637 = \eta \mathbb{E}_\rho [\mathbb{E}_{\pi_\gamma} [\Delta^3(s, a)] - \mathbb{E}_{\pi_\gamma} [\Delta^2(s, a)] \mathbb{E}_{\pi_\gamma} [\Delta(s, a)]]$$

$$1639 \leq 0,$$

1640 where  $\mathbb{E}_\rho$  is the shorthand of  $\mathbb{E}_{s \sim \rho}$ ,  $\mathbb{E}_{\pi_\gamma}$  is the shorthand of  $\mathbb{E}_{a \sim \pi_\gamma}$  and the inequality holds condi-  
 1641 tioned on the event  $\mathcal{E}(\delta)$  due to Lemma 2.15.  $\square$   
 1642

1643 Finally, we have the proposition that adding some bias term  $b : \mathcal{S} \rightarrow \mathbb{R}$  does not affect the resulting  
 1644 policy.

1645 **Proposition G.3.** Let  $b : \mathcal{S} \rightarrow \mathbb{R}$  be some bias function, then for all  $g \in \mathcal{G}$  we have  $J(\pi_g) =$   
 1646  $J(\pi_{g-b})$ , where  $(g-b)(s, a) = g(s, a) - b(s)$ .  
 1647

1648 *Proof.* For any fixed state  $s \in \mathcal{S}$ , we have for any  $a \in \mathcal{A}$  that,

$$1649 \pi_g(a|s) = \frac{\pi^{\text{ref}}(a|s) \exp(\eta g(s, a))}{\sum_{a' \in \mathcal{A}} \pi^{\text{ref}}(a'|s) \exp(\eta g(s, a'))}$$

$$1650 = \frac{\pi^{\text{ref}}(a|s) \exp(\eta g(s, a)) \exp(-\eta b(s))}{\sum_{a' \in \mathcal{A}} \pi^{\text{ref}}(a'|s) \exp(\eta g(s, a')) \exp(-\eta b(s))}$$

$$1651 = \frac{\pi^{\text{ref}}(a|s) \exp(\eta[g(s, a) - b(s)])}{\sum_{a' \in \mathcal{A}} \pi^{\text{ref}}(a'|s) \exp(\eta[g(s, a') - b(s)])}$$

$$1652 = \pi_{g-b}(a|s),$$

1653 which indicates that  $\pi_g = \pi_{g-b}$ . This immediately leads to  $J(\pi_g) = J(\pi_{g-b})$ .  $\square$   
 1654

1655 Now we are ready to prove Theorem F.4.

1656 *Proof of Theorem F.4.* We proceed the proof under the event  $\mathcal{E}(\delta)$ . By Proposition G.3, we know  
 1657 that

$$1658 J(\pi^*) - J(\hat{\pi}) = J(\pi^*) - J(\pi_{\hat{g}})$$

$$1659 = J(\pi^*) - J(\pi_{\hat{g}-b}).$$

1660 Consequently, there exist some  $\gamma \in [0, 1]$  and  $b : \mathcal{S} \rightarrow [-1, 1]$  such that  
 1661

$$1662 J(\pi^*) - J(\hat{\pi}) = J(\pi^*) - J(\pi_{\hat{g}-b})$$

$$1663 \leq \eta \mathbb{E}_{\rho \times \pi_\gamma} [(\hat{g}(s, a) - b(s) - g^*(s, a))^2]$$

$$1664 = \eta G(\gamma),$$

where the inequality holds due to Lemma 2.14. Under event  $\mathcal{E}(\delta)$ , we know that  $\widehat{g}(s, a) - b(s) \leq g^*(s, a)$ . Together with Lemma G.2, we obtain  $G(\gamma) \leq G(0)$ . Therefore, we know that

$$J(\pi^*) - J(\widehat{\pi}) \leq G(0) \quad (\text{G.9})$$

$$\begin{aligned} &= \eta \mathbb{E}_{\rho \times \pi^*} \left[ (\widehat{g}(s, a) - b(s) - g^*(s, a))^2 \right] \\ &\leq 4\eta \left( \mathbb{E}_{\rho \times \pi^*} [\Gamma_n^2(s, a)] \wedge C^{\pi^*} \mathbb{E}_{\rho \times \pi^*} [(\widehat{g}(s, a) - b(s) - g^*(s, a))^2] \right) \\ &= 4\eta \left( \beta^2 \mathbb{E}_{\rho \times \pi^*} [D_{\mathcal{G}}^2((s, a); \pi^{\text{ref}})] \wedge C^{\pi^*} \mathbb{E}_{\rho \times \pi^*} [(\widehat{g}(s, a) - b(s) - g^*(s, a))^2] \right) \\ &= \widetilde{O}(\eta D_{\pi^*}^2 \log \mathcal{G}(\epsilon_c) n^{-1}), \end{aligned} \quad (\text{G.10})$$

where the inequality holds due to the definition of  $\mathcal{E}(\delta)$ . Plugging (G.10) into (G.8), we know that  $J(\pi^*) - J(\widehat{\pi})$  has upper bound  $\widetilde{O}(D_{\pi^*}^2 n^{-1})$ . By Lemma G.1, event  $\mathcal{E}$  with probability at least  $1 - \delta$ , which concludes the proof.  $\square$

## G.2 PROOF OF THEOREM F.6

*Proof of Theorem F.6.* The proof is similar to the proof of Theorem 2.11. Consider the following family of contextual dueling bandit instances with  $S := |\mathcal{S}|, A := |\mathcal{A}| < \infty$  and reward in some function class  $\mathcal{G}$ .

$$\text{CDB} := \{(\mathcal{S}, \mathcal{A}, \rho, r, \pi^{\text{ref}}, \eta) : r \in \mathcal{G}, \rho \in \Delta(\mathcal{S}), \pi^{\text{ref}} \in \Delta(\mathcal{A}|\mathcal{S})\}. \quad (\text{G.11})$$

Fixing any  $S \geq 1, \eta > 4 \log 2$  and  $C^* \in (2, \exp(\eta/4)]$ , we aim to prove that, for any estimator  $\mathcal{D} \mapsto \widehat{\pi} \in \Delta(\mathcal{A}|\mathcal{S})$ , for any  $n \geq 16SC^*$ , there exist some function class  $\mathcal{G}$ , such that  $\exists \text{inst} = (\mathcal{S}, \mathcal{A}, \rho, r, \pi^{\text{ref}}, \eta) \in \text{CDB}$  with single-policy concentrability  $C^{\pi^*} \leq C^*$ , regularization coefficient  $\eta, |\mathcal{S}| = S = \Theta(\log |\mathcal{G}|)$ , and

$$\inf_{\text{inst} \in \text{CDB}} \text{SubOpt}_{\text{RKL}}(\widehat{\pi}; \text{inst}) \gtrsim \min\{\eta SC^* n^{-1}, (SC^*)^{1/2} n^{-1/2}\}. \quad (\text{G.12})$$

Since  $\log |\mathcal{G}| \geq \log \mathcal{N}_{\mathcal{G}}(\epsilon)$  for any  $\epsilon \in (0, 1)$ , the above bound yields the desired result.

We construct the same reward function class as in the proof of Theorem 2.11. In particular, we set  $\mathcal{S} = [S], \mathcal{A} = \{\pm 1\}, \rho = \text{Unif}(\mathcal{S})$ , and the reference policy to be

$$\forall s \in \mathcal{S}, \pi^{\text{ref}}(-1|s) = C^{-1}, \pi^{\text{ref}}(+1|s) = 1 - C^{-1};$$

where  $C = C^*$ . Then the total sub-optimality of any  $\pi \in \Delta(\mathcal{A}|\mathcal{S})$  given any reward function  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is

$$\text{SubOpt}_{f_{\text{div}}}(\pi; r) = \frac{1}{S} \sum_{s=1}^S \text{SubOpt}_{f_{\text{div}}}(\pi(\cdot|s); r(s, \cdot)). \quad (\text{G.13})$$

We further let  $\alpha = \eta^{-1} \log(C - 1) \Leftrightarrow C - 1 = \exp(\eta\alpha)$ . We construct  $2^S$  Bernoulli reward functions, in particular,  $\forall \tau \in \{\pm 1\}^S$ , the mean function  $r_{\tau}$  of the reward (indexed by  $\tau$ ) is defined as

$$r_{\tau}(s, -1) = 0.5 + \tau_s \delta, r_{\tau}(s, +1) = 0.5 - \alpha.$$

Then, following the derivation of (D.12) and (D.13), we know that  $\forall s \in \mathcal{S}, \forall \tau, \tau' \in \{\pm 1\}^S$  with  $\tau \sim_s \tau'$ ,

$$\text{SubOpt}_s(\widehat{\pi}; \tau) + \text{SubOpt}_s(\widehat{\pi}; \tau') \geq \frac{\eta \delta^2}{8} \wedge \frac{3\delta}{10}. \quad (\text{G.14})$$

Let  $P_r$  be the distribution of  $(s, a^1, a^2, y)$  for  $s \sim \rho, a^1, a^2 \stackrel{\text{i.i.d.}}{\sim} \pi^{\text{ref}}(\cdot|s)$  and  $y \sim \text{Bern}(\sigma(r(s, a^1) - r(s, a^2)))$ . Now we set  $\delta = \sqrt{S/n}$  and conclude that for  $\tau \sim \tau'$  with  $\tau_s = -\tau'_s$ ,

$$\begin{aligned} &\text{KL}(P_{r_{\tau}} \| P_{r_{\tau'}}) \\ &= \frac{(C - 1)}{SC^2} \sum_{s', a^1, a^2} \text{KL}(\text{Bern}(\sigma(r_{\tau}(s', a^1) - r(s', a^2))) \| \text{Bern}(\sigma(r_{\tau'}(s', a^1) - r(s', a^2)))) \\ &= \frac{2(C - 1)}{SC^2} \left( \text{KL}(\text{Bern}(\sigma(\alpha + \delta)) \| \text{Bern}(\sigma(\alpha - \delta))) \vee \text{KL}(\text{Bern}(\sigma(\alpha - \delta)) \| \text{Bern}(\sigma(\alpha + \delta))) \right). \end{aligned}$$

1728 Since  $\alpha, \delta \in (0, 1/2)$ , by the fact  $\text{KL}(P\|Q) \leq 2Q_{\min}^{-1} \text{TV}(P\|Q)^2$  (see e.g., Polyanskiy & Wu  
 1729 (2025, Section 7.6)), we know that  
 1730

$$\begin{aligned}
 1731 \text{KL}(P_{r_\tau} \| P_{r_{\tau'}}) &\leq \frac{2(C-1)}{SC^2} \frac{4}{1 + \exp(\alpha + \delta)} \left( \frac{1}{1 + \exp(\alpha - \delta)} - \frac{1}{1 + \exp(\alpha + \delta)} \right)^2 \\
 1732 &\leq \frac{4}{3SC} \frac{\exp(2\alpha)(\exp(\delta) - \exp(-\delta))^2}{(1 + \exp(\alpha - \delta))^4} \\
 1733 &\leq \frac{4e}{3SC} (\exp(\delta) - \exp(-\delta))^2 \\
 1734 &\leq 36S^{-1}C^{-1}\delta^2,
 \end{aligned} \tag{G.15}$$

1739 where the second and third inequality hold due to  $\alpha, \delta \leq 1/2$ , and last inequality follows from  
 1740  $\exp(x) - \exp(-x) \leq 3x$  for  $x \in [0, 1/2]$ . Now we set  $\delta = \sqrt{SC/n} \leq 1/4$ . We substitute (G.14)  
 1741 into Assouad's Lemma (Lemma H.3) and obtain that  
 1742

$$\begin{aligned}
 1743 \inf_{\text{inst} \in \text{CDB}} \text{SubOpt}_{\text{RKL}}(\hat{\pi}; \text{inst}) &\geq \frac{1}{4}S \cdot \frac{1}{S} \cdot \left( \frac{\eta\delta^2}{8} \wedge \frac{3\delta}{10} \right) \cdot \min_{\tau \sim \tau'} \exp \left( -\text{KL}(P_{\mathcal{D}_\tau} \| P_{\mathcal{D}_{\tau'}}) \right) \\
 1744 &= \frac{1}{4} \left( \frac{\eta\delta^2}{8} \wedge \frac{3\delta}{10} \right) \exp \left( -n\text{KL}(P_{r_\tau} \| P_{r_{\mathcal{D}_{\tau'}}}) \right) \\
 1745 &\geq \frac{\exp(-36)}{32} \min\{\eta CSn^{-1}, S^2C^2n^{-2}\},
 \end{aligned}$$

1750 where the  $1/S$  comes from the denominator of (G.13) and the second inequality follows from (G.15).  $\square$   
 1751

### 1753 G.3 PROOF OF THEOREM F.8

1755 *Proof of Theorem F.8.* The proof is similar to the proof of Theorem 3.2. Recall that  $b(\cdot)$  defined  
 1756 in (G.2), we know that  
 1757

$$\begin{aligned}
 1758 \hat{\pi} &= \operatorname{argmax}_{\pi \in \Delta^d} \left\{ \mathbb{E}_{(s,a) \sim \rho \times \pi} [\bar{g}(s, a)] - \eta^{-1} \mathbb{E}_{s \sim \rho} [D_f(\pi \| \pi^{\text{ref}})] \right\} \\
 1759 &= \operatorname{argmax}_{\pi \in \Delta^d} \left\{ \mathbb{E}_{(s,a) \sim \rho \times \pi} [\bar{g}(s, a) - b(s)] - \eta^{-1} \mathbb{E}_{s \sim \rho} [D_f(\pi \| \pi^{\text{ref}})] \right\}.
 \end{aligned}$$

1763 We have the following sub-optimality decomposition

$$\begin{aligned}
 1764 J(\pi^*) - J(\hat{\pi}) &= \mathbb{E}_{s \sim \rho} \left[ \mathbb{E}_{a \sim \pi^*} [g^*(s, a)] - \mathbb{E}_{a \sim \hat{\pi}} [g^*(s, a)] - \eta^{-1} [D_f(\pi^* \| \pi^{\text{ref}}) - D_f(\hat{\pi} \| \pi^{\text{ref}})] \right] \\
 1765 &= \mathbb{E}_{s \sim \rho} [H_s^*(g^*) - H_s^*(\bar{g} - b) - \langle \hat{\pi}, g^* - \bar{g} + b \rangle] \\
 1766 &= \mathbb{E}_{s \sim \rho} [H_s^*(g^*) - H_s^*(\bar{g} - b) - \langle \nabla H_s^*(\bar{g} - b), g^* - \bar{g} + b \rangle] \\
 1767 &= \mathbb{E}_{s \sim \rho} [(g^* - \bar{g} + b)^\top \nabla^2 H_s^*(\bar{g})(g^* - \bar{g} + b)],
 \end{aligned}$$

1771 where  $\tilde{g} = \gamma g^* + (1 - \gamma)\bar{g}$  and  $\gamma \in [0, 1]$ ,  $(\bar{g} - b)(s, a) = \bar{g}(s, a) - b(s)$  and the last equation holds  
 1772 due to Taylor's expansion. Now, for any  $\delta \in (0, 1)$  and  $\epsilon_c > 0$ , with probability at least  $1 - \delta$

$$\begin{aligned}
 1773 J(\pi^*) - J(\hat{\pi}) &= \mathbb{E}_{s \sim \rho} [(g^* - \bar{g} + b)^\top \nabla^2 H_s^*(\tilde{g})(g^* - \bar{g} + b)] \\
 1774 &\leq \alpha^{-1} \eta \mathbb{E}_{s \sim \rho} [(g^* - \bar{g} + b)^\top \operatorname{diag}(\pi^{\text{ref}}(a_1|s), \dots, \pi^{\text{ref}}(a_d|s))(g^* - \bar{g} + b)] \\
 1775 &= \alpha^{-1} \eta \mathbb{E}_{(s,a) \sim \rho \times \pi^{\text{ref}}} [(g^*(s, a) - \bar{g}(s, a) + b(s))^2] \\
 1776 &\leq \alpha^{-1} \eta \left( \frac{128}{3n} \log(2\mathcal{N}_G(\epsilon_c)/\delta) + 18\epsilon_c \right),
 \end{aligned}$$

1778 where the first inequality holds due to Lemma E.4 and last inequality holds due to equation (G.3).  
 1779 Setting  $\epsilon_c = O(n^{-1})$  completes the proof.  $\square$

1782 G.4 PROOF OF THEOREM F.9  
1783

1784 *Proof of Theorem F.9.* We still consider the contextual dueling bandit instance class defined  
1785 in (G.11). We show that given any positive  $\alpha, \eta$ , for any  $n \geq S \cdot \max\{16, \eta^2 \alpha^{-2}\}$ , there exists  
1786  $f : \mathbb{R} \rightarrow \mathbb{R}$  such that  $f$  is  $\alpha$ -strongly convex,  $\log |\mathcal{G}| = \Theta(S)$  and

$$1787 \inf_{\hat{\pi} \in \hat{\Pi}(\mathcal{D})} \sup_{\text{inst} \in \text{CDB}} \text{SubOpt}_{f\text{div}}(\hat{\pi}; \text{inst}) \gtrsim \frac{\eta S}{\alpha n}, \quad (\text{G.16})$$

1790 where  $\mathcal{D} = \{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$  is the offline preference dataset, all (possibly randomized) maps  
1791 from which to  $\Delta(\mathcal{A}|\mathcal{S})$  is denoted by  $\hat{\Pi}(\mathcal{D})$ . Since  $S = \Theta(\log |\mathcal{G}|) \gtrsim \log \mathcal{N}_{\mathcal{G}}(\epsilon_c)$  for all  $\epsilon_c \in (0, 1)$ ,  
1792 we can conclude the theorem.  
1793

1794 Let  $\mathcal{S} = [S]$ ,  $\mathcal{A} = \{\pm 1\}$ ,  $\rho = \text{Unif}(\mathcal{S})$  and  $\pi^{\text{ref}}(\cdot|s) = \text{Unif}(\mathcal{A})$  for any  $s \in \mathcal{S}$ . Then the total  
1795 sub-optimality of any  $\pi \in \Delta(\mathcal{A}|\mathcal{S})$  given any reward function  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is

$$1797 \text{SubOpt}_{f\text{div}}(\pi; r) = \frac{1}{S} \sum_{s=1}^S \text{SubOpt}_{f\text{div}}(\pi(\cdot|s); r(s, \cdot)). \quad (\text{G.17})$$

1800 We still consider the reward function class  $\mathcal{G}$  indexed by  $\{\pm 1\}^S$ . For all  $\tau \in \{\pm 1\}^S$  the reward  
1801 instance “shaped” by  $\tau$  is  
1802

$$1803 r_{\tau}(s, a) = \frac{1}{2} + a\tau_s \cdot \sqrt{\frac{S}{n}}, \quad (\text{G.18})$$

1804 where  $a\tau_s = \pm 1$  because  $a \in \mathcal{A} = \{\pm 1\}$ . We thereby refer  $\tau \sim \tau'$  to any pair in  $\{\pm 1\}^S$  that differs  
1805 only in one coordinate.  $\forall \tau, \tau' \in \{\pm 1\}^S$ , if  $\tau \sim \tau'$ , then suppose  $\tau_s = -\tau'_s$ , we have  
1806

$$1807 \text{SubOpt}_{f\text{div}}(\pi(\cdot|s); r_{\tau}(s, \cdot)) + \text{SubOpt}_{f\text{div}}(\pi(\cdot|s); r_{\tau'}(s, \cdot)) \geq \frac{\eta S}{\alpha n}, \quad (\text{G.19})$$

1808 where the inequality follows from exactly the same calculation in equation (E.5) by setting  $f(x) =$   
1809  $\alpha(x - 1)^2/2$ .<sup>7</sup> Let  $P_r$  be the distribution of  $(s, a^1, a^2, y)$  for  $s \sim \rho$ ,  $a^1, a^2 \stackrel{\text{i.i.d.}}{\sim} \pi^{\text{ref}}(\cdot|s)$  and  
1810  $y \sim \text{Bern}(\sigma(r(s, a^1) - r(s, a^2)))$ . Then we denote  $\delta = \sqrt{S/n}$  and conclude that for  $\tau \sim \tau'$  with  
1811  $\tau_s = -\tau'_s$ ,  
1812

$$1813 \text{KL}(P_{r_{\tau}} \| P_{r_{\tau'}}) = \frac{1}{SA^2} \sum_{s', a^1, a^2} \text{KL}(\text{Bern}(\sigma(r_{\tau}(s', a^1) - r(s', a^2))) \| \text{Bern}(\sigma(r_{\tau'}(s', a^1) - r(s', a^2))))$$

$$1814 = \frac{1}{4S} \left( \text{KL}(\text{Bern}(\sigma(2\delta)) \| \text{Bern}(\sigma(-2\delta))) + \text{KL}(\text{Bern}(\sigma(-2\delta)) \| \text{Bern}(\sigma(2\delta))) \right)$$

$$1815 \leq \frac{1}{4S} \left( (\exp(-2\delta) - 1)^2 + (\exp(2\delta) - 1)^2 \right)$$

$$1816 \leq \frac{1}{2S} (\exp(2\delta) - 1)^2 \leq \frac{36\delta^2}{2S} = \frac{18}{n}, \quad (\text{G.20})$$

1817 where the last inequality follows from  $\exp(x) - 1 \leq 3x$  for  $x \in [0, 0.5]$  and  $\delta = \sqrt{S/n} \leq 0.25$  by  
1818 assumption. Therefore, we substitute (G.19) into Assouad’s Lemma (Lemma H.3) to obtain  
1819

$$1820 \text{LHS of (G.16)} \geq \frac{1}{S} \cdot S \cdot \frac{\eta S}{\alpha n} \cdot \frac{1}{4} \cdot \min_{\tau \sim \tau'} \exp \left( -\text{KL}(P_{r_{\tau}} \| P_{r_{\tau'}}) \right)$$

$$1821 = 0.25 \cdot \frac{\eta S}{\alpha n} \cdot \exp \left( -n \text{KL}(P_{r_{\tau}} \| P_{r_{\tau'}}) \right) \geq \frac{\eta S}{\alpha n} \cdot \frac{1}{3} \cdot \exp(-18) \gtrsim \frac{\eta S}{\alpha n}, \quad (\text{G.21})$$

1822 where the  $1/S$  comes from the denominator of (G.17) and the second inequality follows from (G.20).  
1823  $\square$

1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
2212  
2213  
2214  
2215  
2216  
2217  
2218  
2219  
2220  
2221  
2222  
2223  
2224  
2225  
2226  
2227  
2228  
2229  
2230  
2231  
2232  
2233  
2234  
2235  
2236  
2237  
2238  
2239  
2240  
2241  
2242  
2243  
2244  
2245  
2246  
2247  
2248  
2249  
2250  
2251  
2252  
2253  
2254  
2255  
2256  
2257  
2258  
2259  
2260  
2261  
2262  
2263  
2264  
2265  
2266  
2267  
2268  
2269  
2270  
2271  
2272  
2273  
2274  
2275  
2276  
2277  
2278  
2279  
2280  
2281  
2282  
2283  
2284  
2285  
2286  
2287  
2288  
2289  
2290  
2291  
2292  
2293  
2294  
2295  
2296  
2297  
2298  
2299  
2299  
2300  
2301  
2302  
2303  
2304  
2305  
2306  
2307  
2308  
2309  
2310  
2311  
2312  
2313  
2314  
2315  
2316  
2317  
2318  
2319  
2320  
2321  
2322  
2323  
2324  
2325  
2326  
2327  
2328  
2329  
2330  
2331  
2332  
2333  
2334  
2335  
2336  
2337  
2338  
2339  
2340  
2341  
2342  
2343  
2344  
2345  
2346  
2347  
2348  
2349  
2350  
2351  
2352  
2353  
2354  
2355  
2356  
2357  
2358  
2359  
2360  
2361  
2362  
2363  
2364  
2365  
2366  
2367  
2368  
2369  
2370  
2371  
2372  
2373  
2374  
2375  
2376  
2377  
2378  
2379  
2380  
2381  
2382  
2383  
2384  
2385  
2386  
2387  
2388  
2389  
2390  
2391  
2392  
2393  
2394  
2395  
2396  
2397  
2398  
2399  
2400  
2401  
2402  
2403  
2404  
2405  
2406  
2407  
2408  
2409  
2410  
2411  
2412  
2413  
2414  
2415  
2416  
2417  
2418  
2419  
2420  
2421  
2422  
2423  
2424  
2425  
2426  
2427  
2428  
2429  
2430  
2431  
2432  
2433  
2434  
2435  
2436  
2437  
2438  
2439  
2440  
2441  
2442  
2443  
2444  
2445  
2446  
2447  
2448  
2449  
2450  
2451  
2452  
2453  
2454  
2455  
2456  
2457  
2458  
2459  
2460  
2461  
2462  
2463  
2464  
2465  
2466  
2467  
2468  
2469  
2470  
2471  
2472  
2473  
2474  
2475  
2476  
2477  
2478  
2479  
2480  
2481  
2482  
2483  
2484  
2485  
2486  
2487  
2488  
2489  
2490  
2491  
2492  
2493  
2494  
2495  
2496  
2497  
2498  
2499  
2499  
2500  
2501  
2502  
2503  
2504  
2505  
2506  
2507  
2508  
2509  
25010  
25011  
25012  
25013  
25014  
25015  
25016  
25017  
25018  
25019  
25020  
25021  
25022  
25023  
25024  
25025  
25026  
25027  
25028  
25029  
25030  
25031  
25032  
25033  
25034  
25035  
25036  
25037  
25038  
25039  
25040  
25041  
25042  
25043  
25044  
25045  
25046  
25047  
25048  
25049  
25050  
25051  
25052  
25053  
25054  
25055  
25056  
25057  
25058  
25059  
25060  
25061  
25062  
25063  
25064  
25065  
25066  
25067  
25068  
25069  
25070  
25071  
25072  
25073  
25074  
25075  
25076  
25077  
25078  
25079  
25080  
25081  
25082  
25083  
25084  
25085  
25086  
25087  
25088  
25089  
25090  
25091  
25092  
25093  
25094  
25095  
25096  
25097  
25098  
25099  
25099  
25100  
25101  
25102  
25103  
25104  
25105  
25106  
25107  
25108  
25109  
25110  
25111  
25112  
25113  
25114  
25115  
25116  
25117  
25118  
25119  
25120  
25121  
25122  
25123  
25124  
25125  
25126  
25127  
25128  
25129  
25130  
25131  
25132  
25133  
25134  
25135  
25136  
25137  
25138  
25139  
25140  
25141  
25142  
25143  
25144  
25145  
25146  
25147  
25148  
25149  
25150  
25151  
25152  
25153  
25154  
25155  
25156  
25157  
25158  
25159  
25160  
25161  
25162  
25163  
25164  
25165  
25166  
25167  
25168  
25169  
25170  
25171  
25172  
25173  
25174  
25175  
25176  
25177  
25178  
25179  
25180  
25181  
25182  
25183  
25184  
25185  
25186  
25187  
25188  
25189  
25190  
25191  
25192  
25193  
25194  
25195  
25196  
25197  
25198  
25199  
25199  
25200  
25201  
25202  
25203  
25204  
25205  
25206  
25207  
25208  
25209  
25210  
25211  
25212  
25213  
25214  
25215  
25216  
25217  
25218  
25219  
25220  
25221  
25222  
25223  
25224  
25225  
25226  
25227  
25228  
25229  
25230  
25231  
25232  
25233  
25234  
25235  
25236  
25237  
25238  
25239  
25240  
25241  
25242  
25243  
25244  
25245  
25246  
25247  
25248  
25249  
25250  
25251  
25252  
25253  
25254  
25255  
25256  
25257  
25258  
25259  
25260  
25261  
25262  
25263  
25264  
25265  
25266  
25267  
25268  
25269  
25270  
25271  
25272  
25273  
25274  
25275  
25276  
25277  
25278  
25279  
25280  
25281  
25282  
25283  
25284  
25285  
25286  
25287  
25288  
25289  
25290  
25291  
25292  
25293  
25294  
25295  
25296  
25297  
25298  
25299  
25299  
25300  
25301  
25302  
25303  
25304  
25305  
25306  
25307  
25308  
25309  
25310  
25311  
25312  
25313  
25314  
25315  
25316  
25317  
25318  
25319  
25320  
25321  
25322  
25323  
25324  
25325  
25326  
25327  
25328  
25329  
25330  
25331  
25332  
25333  
25334  
25335  
25336  
25337  
25338  
25339  
25340  
25341  
25342  
25343  
25344  
25345  
25346  
25347  
25348  
25349  
25350  
25351  
25352  
25353  
25354  
25355  
25356  
25357  
25358  
25359  
25360  
25361  
25362  
25363  
25364  
25365  
25366  
25367  
25368  
25369  
25370  
25371  
25372  
25373  
25374  
25375  
25376  
25377  
25378  
25379  
25380  
25381  
25382  
25383  
25384  
25385  
25386  
25387  
25388  
25389  
25390  
25391  
25392  
25393  
25394  
25395  
25396  
25397  
25398  
25399  
25399  
25400  
25401  
25402  
25403  
25404  
25405  
25406  
25407  
25408  
25409  
25410  
25411  
25412  
25413  
25414  
25415  
25416  
25417  
25418  
25419  
25420  
25421  
25422  
25423  
25424  
25425  
25426  
25427  
25428  
25429  
25430  
25431  
25432  
25433  
25434  
25435  
25436  
25437  
25438  
25439  
25440  
25441  
25442  
25443  
25444  
25445  
25446  
25447  
25448  
25449  
25450  
25451  
25452  
25453  
25454  
25455  
25456  
25457  
25458  
25459  
25460  
25461  
25462  
25463  
25464  
25465  
25466  
25467  
25468  
25469  
25470  
25471  
25472  
25473  
25474  
25475  
25476  
25477  
25478  
25479  
25480  
25481  
25482  
25483  
25484  
25485  
25486  
25487  
25488  
25489  
25490  
25491  
25492  
25493  
25494  
25495  
25496  
25497  
25498  
25499  
25499  
25500  
25501  
25502  
25503  
25504  
25505  
25506  
25507  
25508  
25509  
25510  
25511  
25512  
25513  
25514  
25515  
25516  
25517  
25518  
25519  
25520  
25521  
25522  
25523  
25524  
25525  
25526  
25527  
25528  
25529  
25530  
25531  
25532  
25533  
25534  
25535  
25536  
25537  
25538  
25539  
25540  
25541  
25542  
25543  
25544  
25545  
25546  
25547  
25548  
25549  
25550  
25551  
25552  
25553  
25554  
25555  
25556  
25557  
25558  
25559  
25560  
25561  
25562  
25563  
25564  
25565  
25566  
25567  
25568  
25569  
25570  
25571  
25572  
25573  
25574  
25575  
25576  
25577  
25578  
25579  
25580  
25581  
25582  
25583  
25584  
25585  
25586  
25587  
25588  
25589  
25590  
25591  
25592  
25593  
25594  
25595  
25596  
25597  
25598  
25599  
25599  
25600  
25601  
25602  
25603  
25604  
25605  
25606  
25607  
25608  
25609  
25610  
25611  
25612  
25613  
25614  
25615  
25616  
25617  
25618  
25619  
25620  
25621  
25622  
25623  
25624  
25625  
25626  
25627  
25628  
25629  
25630  
25631  
25632  
25633  
25634  
25635  
25636  
25637  
25638  
25639  
25640  
25641  
25642  
25643  
25644  
25645  
25646  
25647  
25648  
25649  
25650  
25651  
25652  
25653  
25654  
25655  
25656  
25657  
25658  
25659  
25660  
25661  
25662  
25663  
25664  
25665  
25666  
25667  
25668  
25669  
25670  
25

1836 

## H AUXILIARY LEMMAS

1837

1838 **Lemma H.1** (Zhao et al. 2024, Lemma D.4). Consider a offline dataset  $\{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$  generated  
1839 from the product of the context distribution  $\rho \in \Delta(\mathcal{S})$ , policy  $\pi \in \Delta(\mathcal{A}|\mathcal{S})$ , and the Bradley-  
1840 Terry Model defined in Appendix F.1. Suppose  $\bar{g}$  is the result of MLE estimation of Algorithm 3,  
1841 and we further define  $b(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [\bar{g}(s, a) - g^*(s, a)]$ , then with probability at least  $1 - 2\delta$ , we  
1842 have

1843 
$$\mathbb{E}_{s, a \sim \rho \times \pi} [(\bar{g}(s, a) - g^*(s, a) - b(s))^2] \leq O\left(\frac{1}{n} \log(\mathcal{N}_G(\epsilon_c)/\delta) + \epsilon_c\right).$$
1844

1845 Lemmas H.2 and H.3 are two standard reductions (Le Cam, 1973; Yu, 1997; Polyanskiy & Wu,  
1846 2025). See, e.g., Chen et al. (2024, Section 3) for a general proof.

1847 **Lemma H.2** (Fano’s inequality). Fix any  $\mathcal{R} := \{r_1, \dots, r_S\}$  and policy class  $\Pi$ , let  $L : \Pi \times \mathcal{R} \rightarrow \mathbb{R}_+$  be some loss function. Suppose there exist some constant  $c > 0$  such that the following condition holds:

1848 
$$\min_{i \neq j} \min_{\pi \in \Pi} L(\pi, r_i) + L(\pi, r_j) \geq c.$$
1849

1850 Then we have

1851 
$$\inf_{\pi \in \Pi} \sup_{r \in \mathcal{R}} L(\pi, r) \geq \frac{c}{2} \left(1 - \frac{\max_{i \neq j} \text{KL}(P_{r_i} \| P_{r_j}) + \log 2}{\log S}\right),$$
1852

1853 where  $P_r$  is the distribution of dataset given model  $r \in \mathcal{R}$ .

1854 **Lemma H.3** (Assouad’s Lemma). Let  $\mathcal{R}$  be the set of instances,  $\Pi$  be the set of estimators,  $\Theta := \{\pm 1\}^S$  for some  $S > 0$ , and  $\{L_j\}_{j=1}^S$  be  $S$  functions from  $\Pi \times \mathcal{R}$  to  $\mathbb{R}_+$ . Suppose  $\{r_\theta\}_{\theta \in \Theta} \subset \mathcal{R}$  and the loss function is

1855 
$$L(\pi, r) := \sum_{j=1}^S L_j(\pi, r), \forall (\pi, r) \in \Pi \times \mathcal{R}.$$
1856

1857 We denote  $\theta \sim_j \theta'$  if they differ only in the  $j$ -th coordinate. Further assume that

1858 
$$\theta \sim_j \theta' \Rightarrow \inf_{\pi \in \Pi} L_j(\pi, r_\theta) + L_j(\pi, r_{\theta'}) \geq c \tag{H.1}$$
1859

1860 for some  $c > 0$ , then

1861 
$$\inf_{\pi \in \Pi} \sup_{r \in \mathcal{R}} L(\pi, r) \geq S \cdot \frac{c}{4} \min_{\exists j: \theta \sim_j \theta'} \exp\left(-\text{KL}(P_{r_\theta} \| P_{r_{\theta'}})\right),$$
1862

1863 where  $P_r$  denotes the distribution of the dataset given  $r \in \mathcal{R}$ .

1864 The following Lemma H.4 is due to Gilbert (1952); Varshamov (1957), which is a classical result in  
1865 coding theory.

1866 **Lemma H.4.** Suppose  $\Sigma$  is a set of characters with  $|\Sigma| = q$  where  $q \geq 2$  is a prime power and  $N > 0$  is some natural number. Then there exists a subset  $\mathcal{V}$  of  $\Sigma^N$  such that (1) for any  $v, v' \in \mathcal{V}, v \neq v_j$ , one has  $d_H(v, v') \geq N/2$  and (2)  $\log_q |\mathcal{V}| \geq H_q(1/2) = \Theta(1)$ , where  $d_H$  is the Hamming distance and the entropy function  $H$  is given by

1867 
$$H_q(x) = x \frac{\log(q-1)}{\log q} - x \frac{\log x}{\log q} - (1-x) \frac{\log(1-x)}{\log q}.$$
1868

1869 For example, when  $q = 2$ , this means that there exists a subset  $\mathcal{V}$  of  $\{-1, 1\}^S$  such that (1)  $|\mathcal{V}| \geq \exp(S/8)$  and (2) for any  $v, v' \in \mathcal{V}, v \neq v_j$ , one has  $\|v - v'\|_1 \geq S/2$ .

1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889