

BAYESIAN OPTIMISATION IS PROBABILISTIC NUMERICS

Michael A Osborne, @maosbot

**Global optimisation is proper
optimisation.**

Exploitation

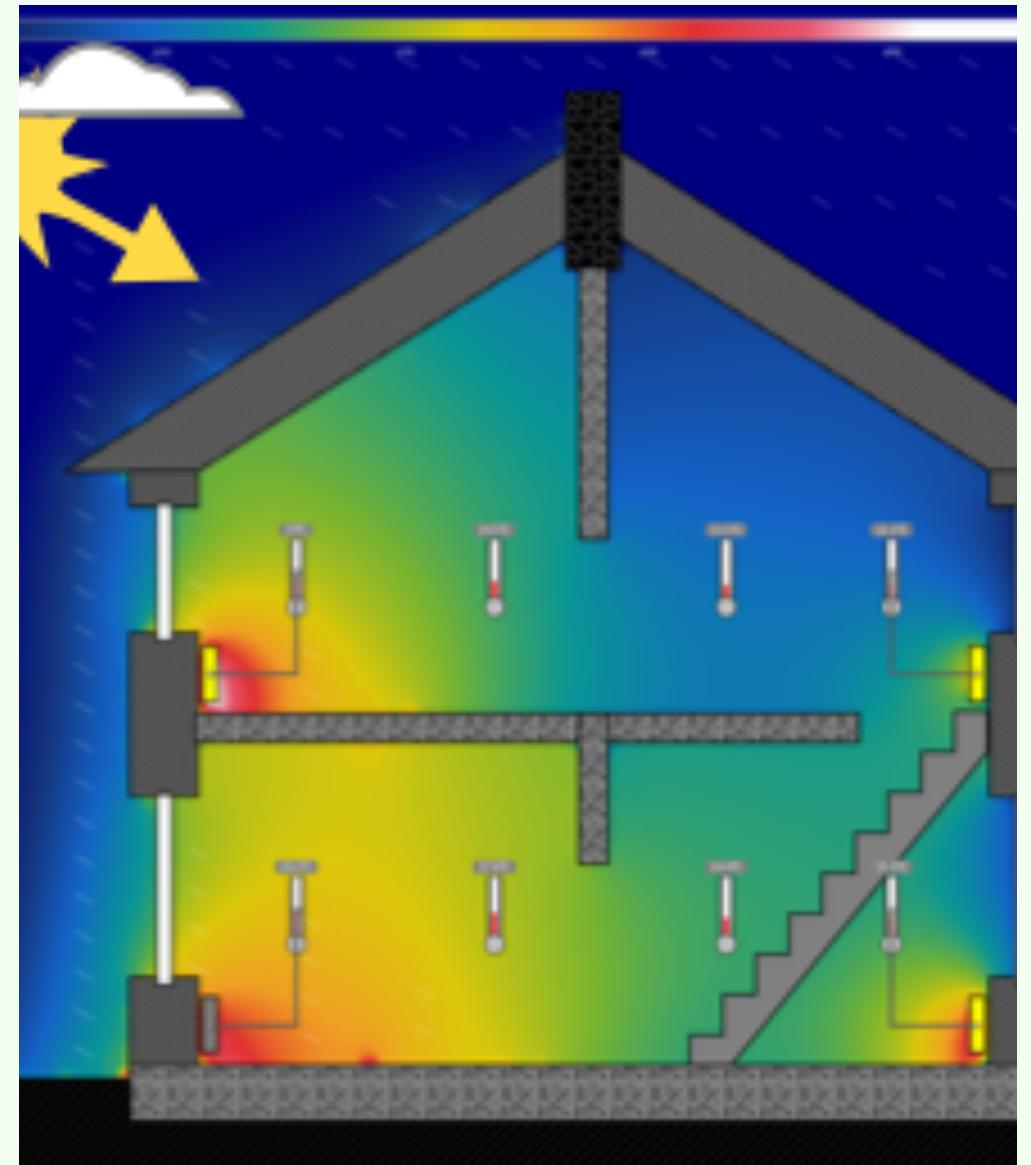
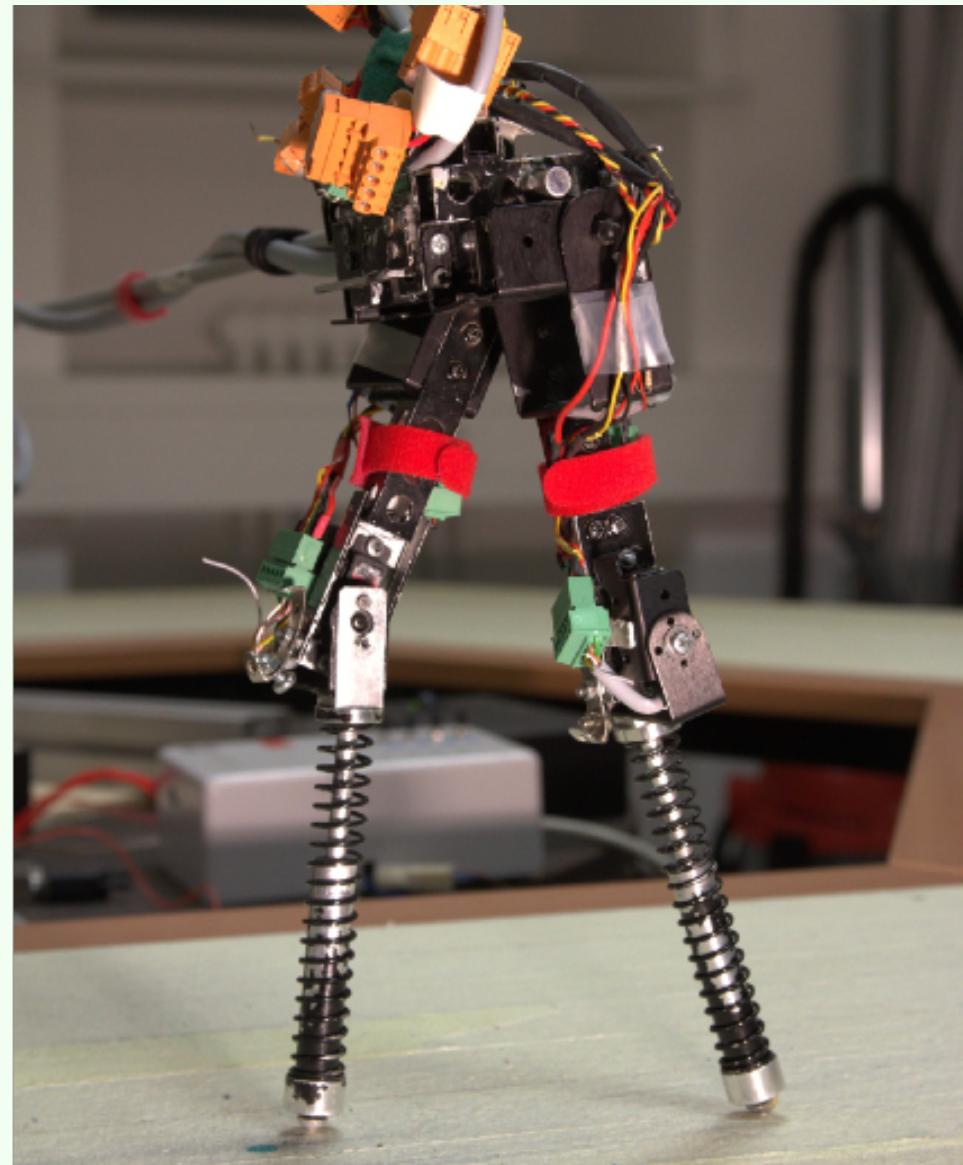
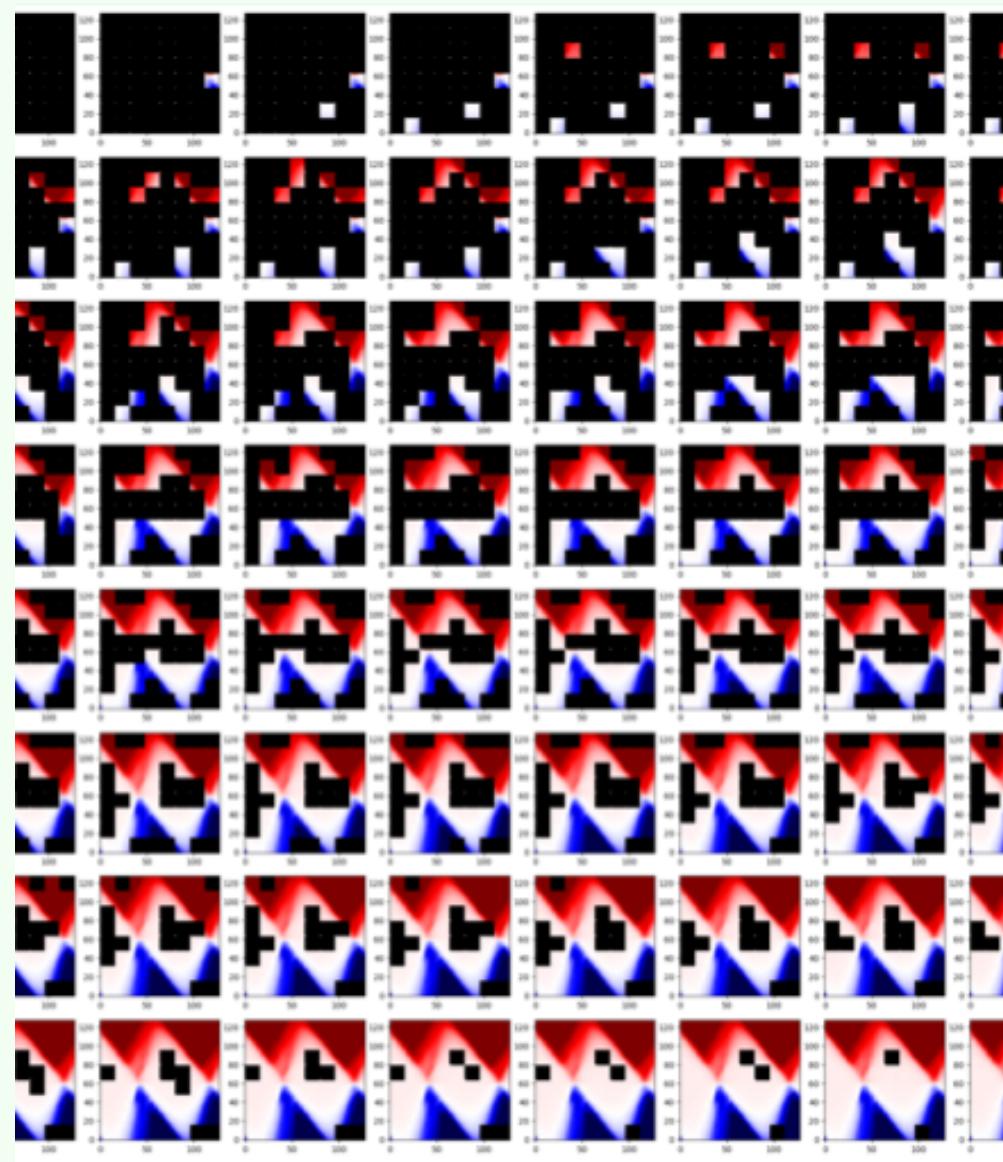
Exploration



Relative to local optimisation,
global optimisation:

1. is less amenable to **theory**;
2. requires higher **overhead**; and
3. overhead costs scale more poorly in **dimension**.

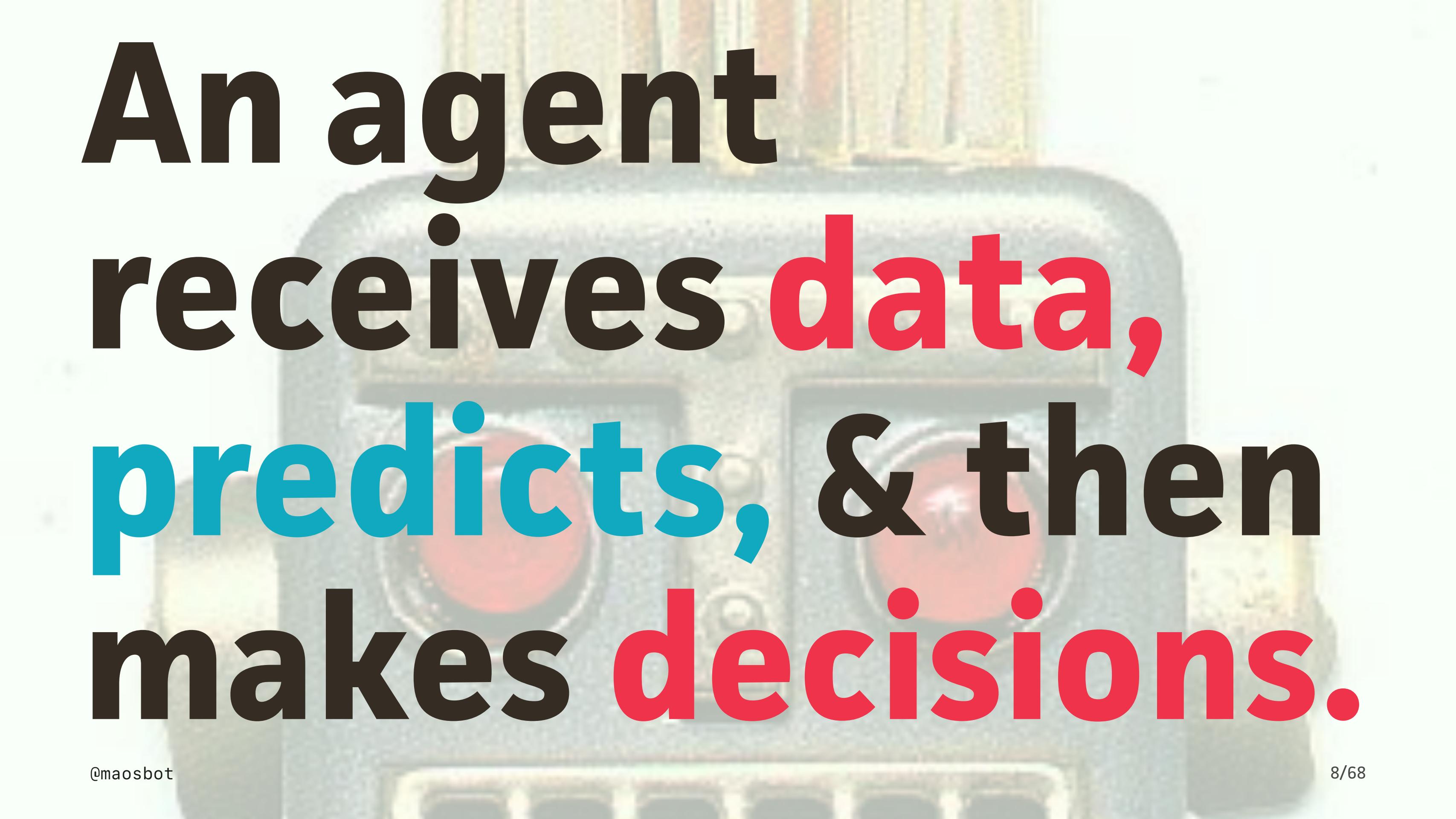
Global optimisation is widely used.



```
9 import numpy as np
10 import platform
11 import subprocess
12 import nlopt
13 from sklearn.utils import check_random_state
14 from scipy.stats import beta, norm
15
16 class RobotArm:
17     def __init__(self):
18         self.name = 'RobotArm Simulator'
19         self.state = np.zeros(3)
20
21     def abs_pos(self, jt_angle):
22         assert jt_angle.ndim == 1, 'jt_angle has to be one dimensional'
23         assert len(jt_angle) == 3, 'jt_angle has to have 3 inputs'
24
25         if platform.system() == 'Windows':
26             args = str('./robot_arm') + str(jt_angle[0]) + '+' + str(jt_angle[1])
27             proc = subprocess.Popen(args, shell=True, stdout=subprocess.PIPE)
28
29         elif platform.system() == 'Darwin':
30             args = str('robot_arm.exe') + str(jt_angle[0]) + '+' + str(jt_angle[1])
31             proc = subprocess.Popen(args, stdout=subprocess.PIPE)
32
33         output = proc.stdout
34         for line in output:
35             output = line
36         proc.kill()
37
38         return np.array([float(out) for out in output.split()])
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
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
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
279
280
281
282
283
284
285
286
287
288
289
289
290
291
292
293
294
295
296
297
298
299
299
300
301
302
303
304
305
306
307
308
309
309
310
311
312
313
314
315
316
317
318
319
319
320
321
322
323
324
325
326
327
328
329
329
330
331
332
333
334
335
336
337
338
339
339
340
341
342
343
344
345
346
347
348
349
349
350
351
352
353
354
355
356
357
358
359
359
360
361
362
363
364
365
366
367
368
369
369
370
371
372
373
374
375
376
377
378
379
379
380
381
382
383
384
385
386
387
388
389
389
390
391
392
393
394
395
396
397
398
399
399
400
401
402
403
404
405
406
407
408
409
409
410
411
412
413
414
415
416
417
418
419
419
420
421
422
423
424
425
426
427
428
429
429
430
431
432
433
434
435
436
437
438
439
439
440
441
442
443
444
445
446
447
448
449
449
450
451
452
453
454
455
456
457
458
459
459
460
461
462
463
464
465
466
467
468
469
469
470
471
472
473
474
475
476
477
478
479
479
480
481
482
483
484
485
486
487
488
489
489
490
491
492
493
494
495
496
497
498
499
499
500
501
502
503
504
505
506
507
508
509
509
510
511
512
513
514
515
516
517
518
519
519
520
521
522
523
524
525
526
527
528
529
529
530
531
532
533
534
535
536
537
538
539
539
540
541
542
543
544
545
546
547
548
549
549
550
551
552
553
554
555
556
557
558
559
559
560
561
562
563
564
565
566
567
568
569
569
570
571
572
573
574
575
576
577
578
579
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
599
600
601
602
603
604
605
606
607
608
609
609
610
611
612
613
614
615
616
617
618
619
619
620
621
622
623
624
625
626
627
628
629
629
630
631
632
633
634
635
636
637
638
639
639
640
641
642
643
644
645
646
647
648
649
649
650
651
652
653
654
655
656
657
658
659
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
679
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
697
698
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
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
778
779
779
780
781
782
783
784
785
786
787
788
788
789
789
790
791
792
793
794
795
796
797
797
798
799
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
838
839
839
840
841
842
843
844
845
846
847
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
868
869
869
870
871
872
873
874
875
876
877
877
878
878
879
879
880
881
882
883
884
885
886
887
887
888
888
889
889
890
891
892
893
894
895
896
896
897
897
898
898
899
899
900
901
902
903
904
905
906
907
908
909
909
910
911
912
913
914
915
916
917
917
918
918
919
919
920
921
922
923
924
925
926
927
927
928
928
929
929
930
931
932
933
934
935
936
937
937
938
938
939
939
940
941
942
943
944
945
946
946
947
947
948
948
949
949
950
951
952
953
954
955
956
956
957
957
958
958
959
959
960
961
962
963
964
965
965
966
966
967
967
968
968
969
969
970
971
972
973
974
975
975
976
976
977
977
978
978
979
979
980
981
982
983
984
985
985
986
986
987
987
988
988
989
989
990
991
992
993
994
995
996
997
998
999
```

Machine learning treats
algorithms as agents.

Probabilistic numerics treats
numeric algorithms as agents.



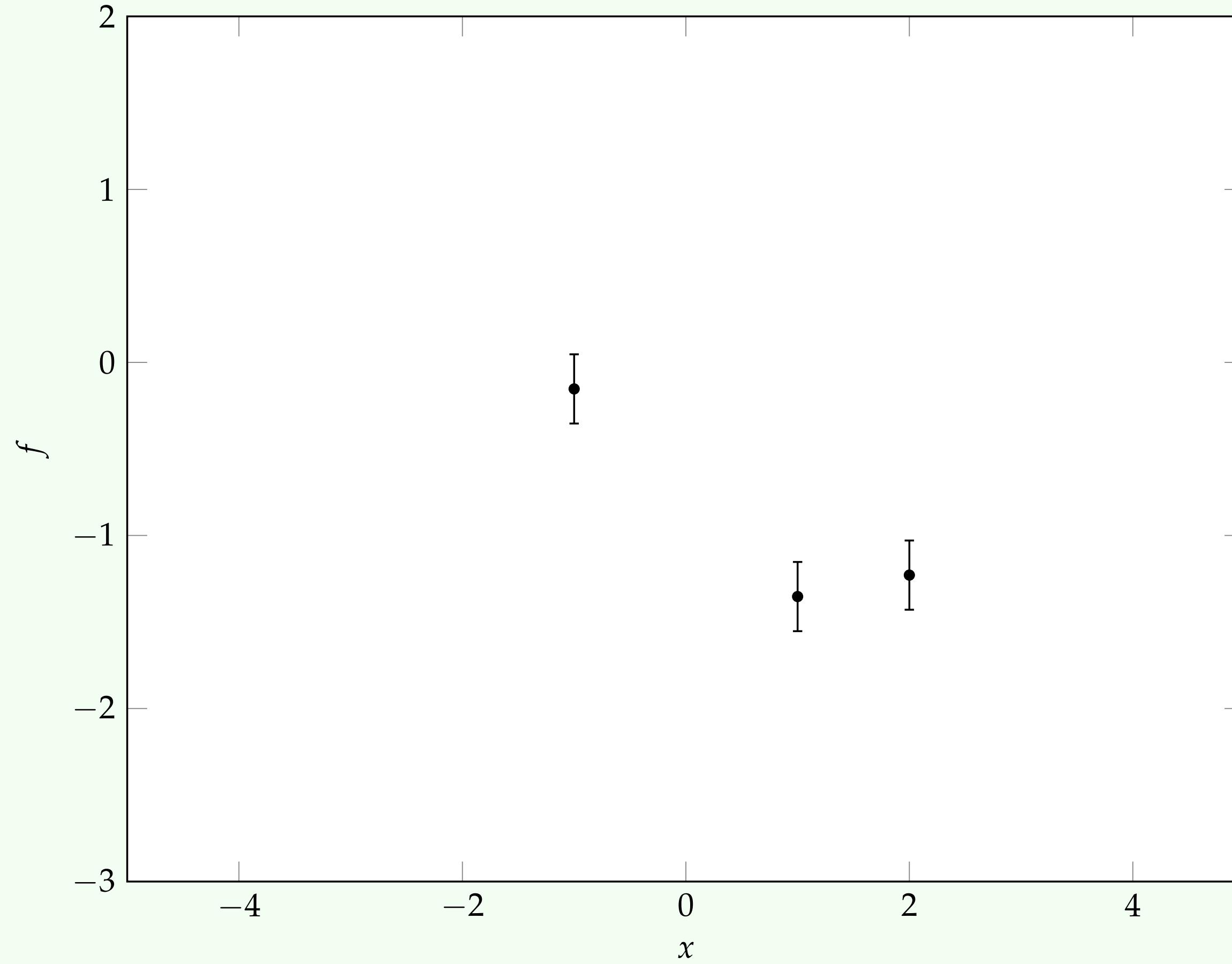
An agent receives data, predicts, & then makes decisions.

In global optimisation:

data = ?;

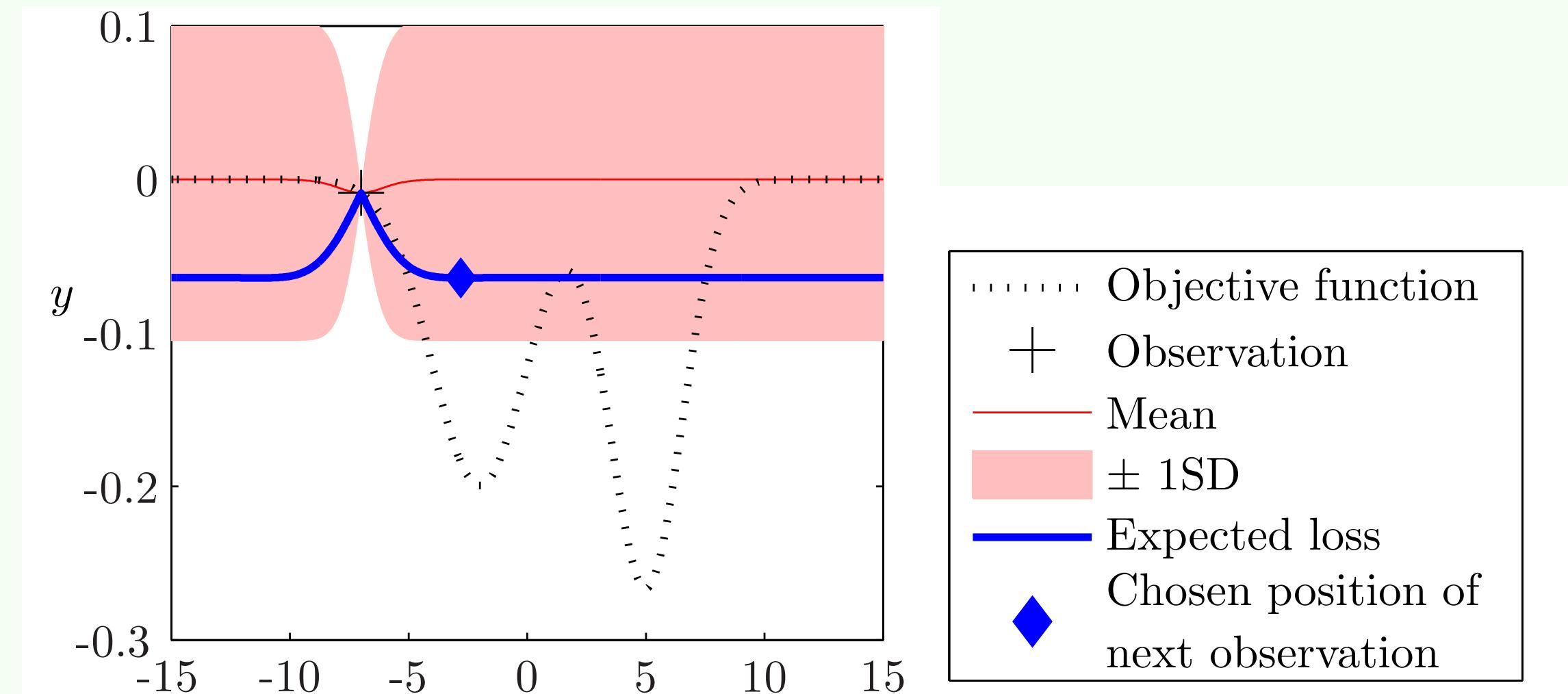
predictand = ?; &

decisions = ?..



In global optimisation:
data = evaluations;
predictand = minimiser; &
decisions = locations.

Bayesian optimisation is probabilistic numerics for global optimisation.





An agent is defined by
its prior and
loss function.

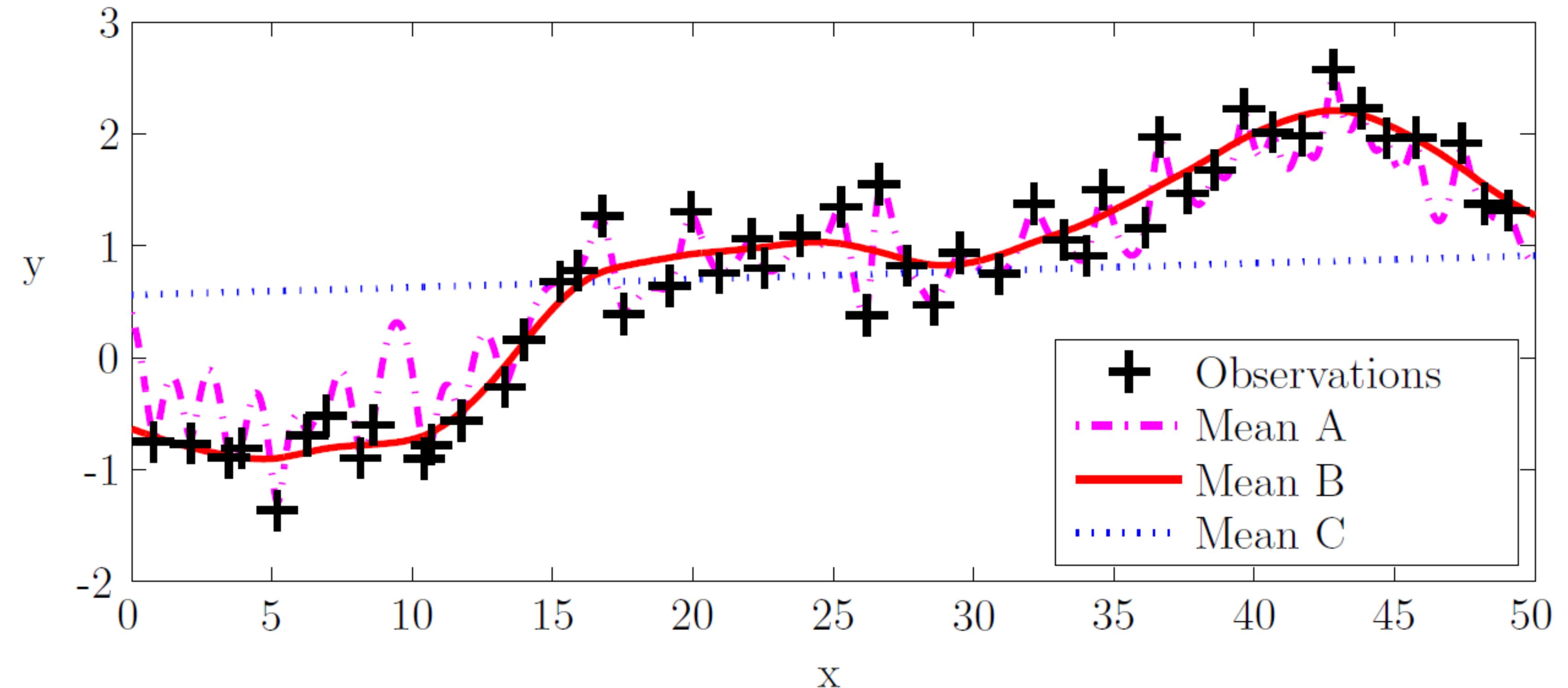
The **surrogate** is the **prior** for the
objective: options include

Gaussian processes,

random forests,

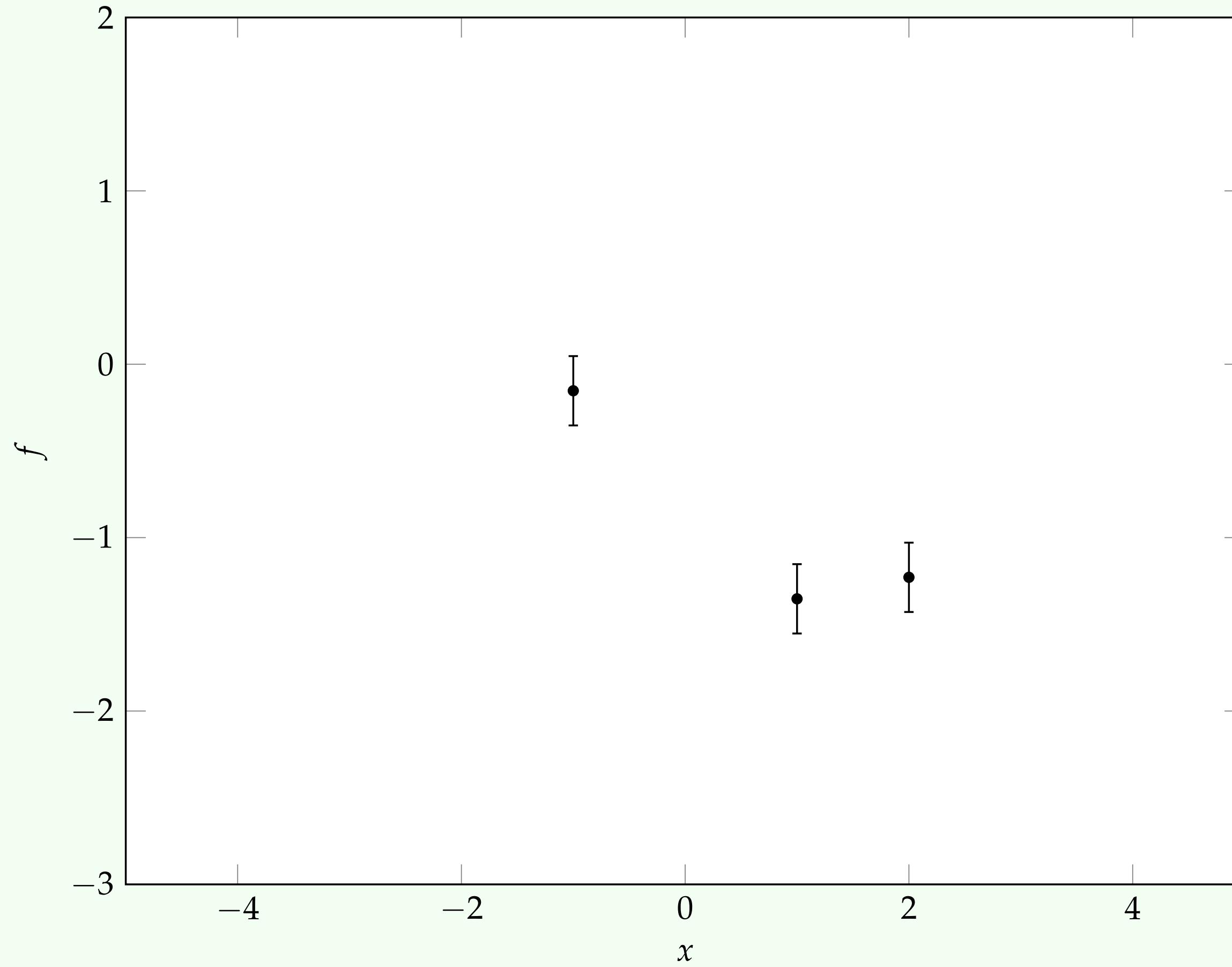
**tree-structured Parzen (density)
estimators and**

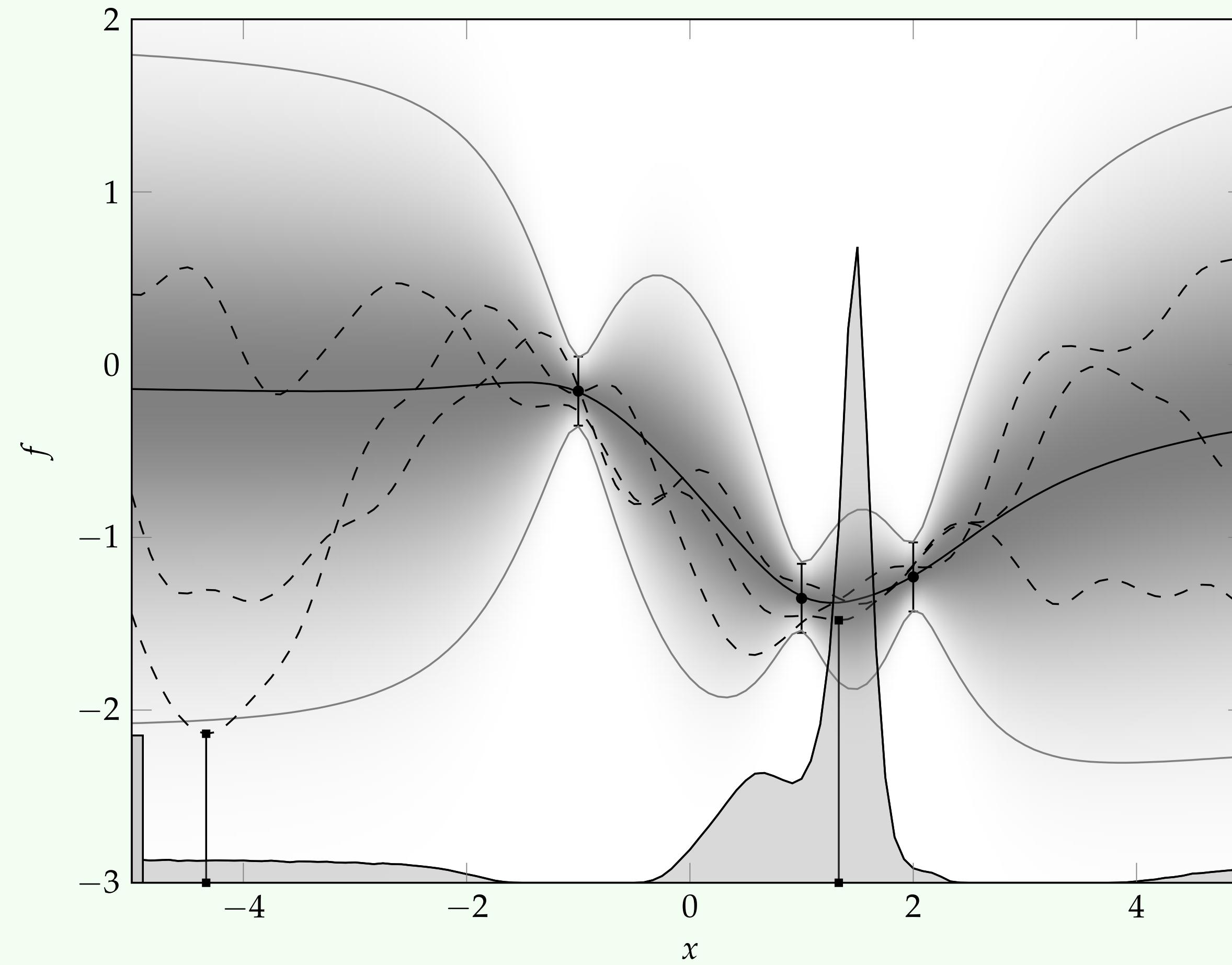
Bayesian neural networks.



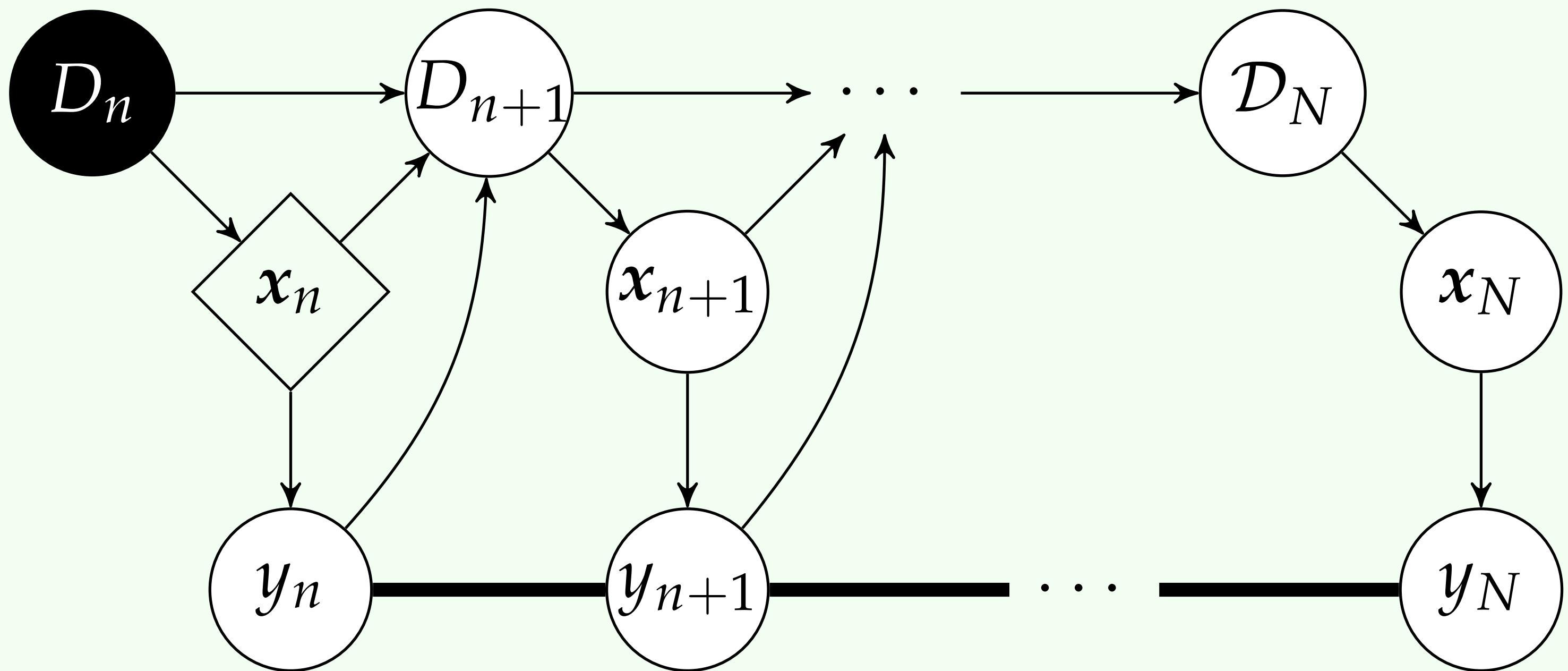
**TO IMPROVE
OPTIMISATION,
IMPROVE YOUR
SURROGATE.**







**Improving calibration is as
important as improving
accuracy.**





What should we pick as the
loss function
for optimisation?

The loss for optimisation could be:

1. the lowest evaluation **(value)**; or
2. the uncertainty in the minimiser **(location-information)**; or
3. the uncertainty in the minimum **(value-information)**.

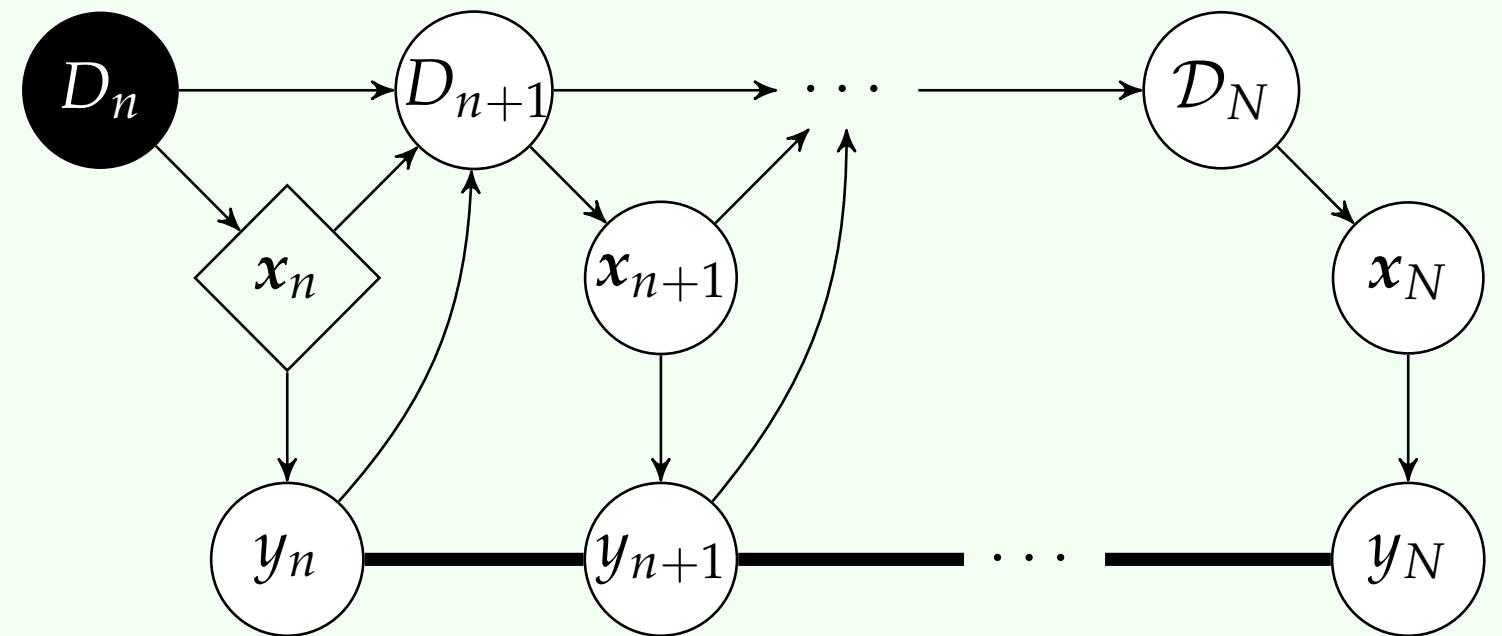
1. Value: $\lambda_{\text{VL}} := y_N$.

2. Location-information:

$$\lambda_{\text{LIL}} := \mathbb{H}(\boldsymbol{x}_* \mid \boldsymbol{x}_N, y_N, \mathcal{D}_N).$$

2. Value-information:

$$\lambda_{\text{VIL}} := \mathbb{H}(y_* \mid \boldsymbol{x}_N, y_N, \mathcal{D}_N).$$

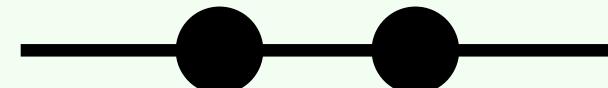


The minimiser is \boldsymbol{x}_* and the minimum y_* .

An acquisition function
is an expected
loss function.

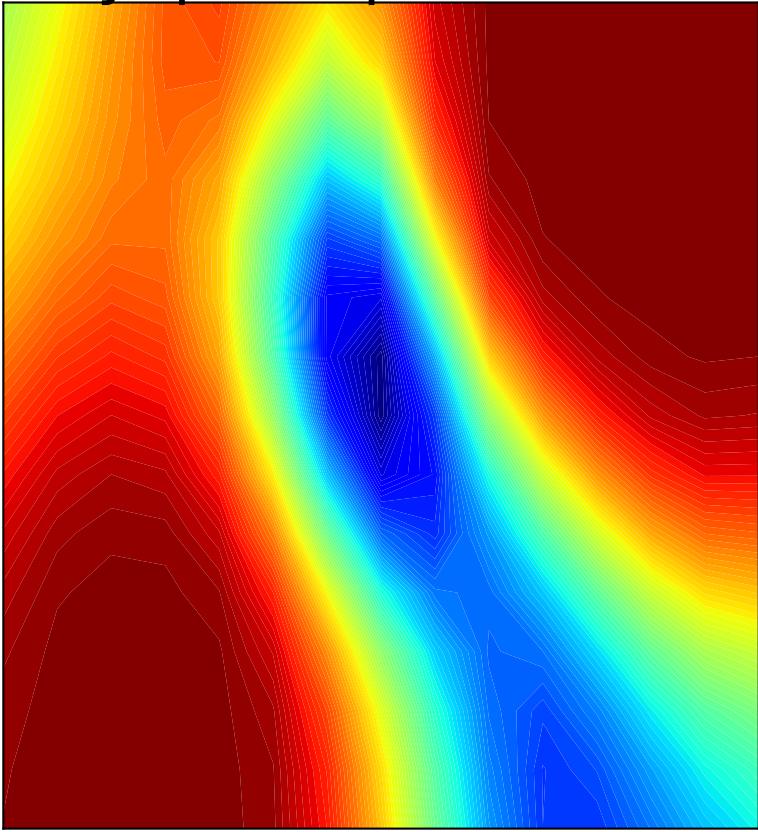


Myopia can lead to insufficient exploration.

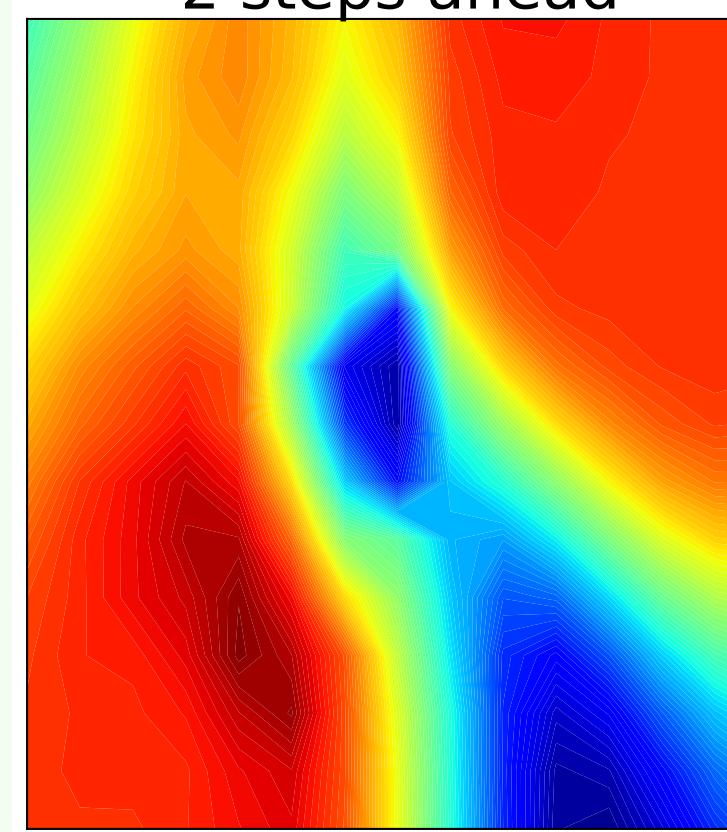


On the other hand, any flaws of a surrogate are magnified by non-myopia.

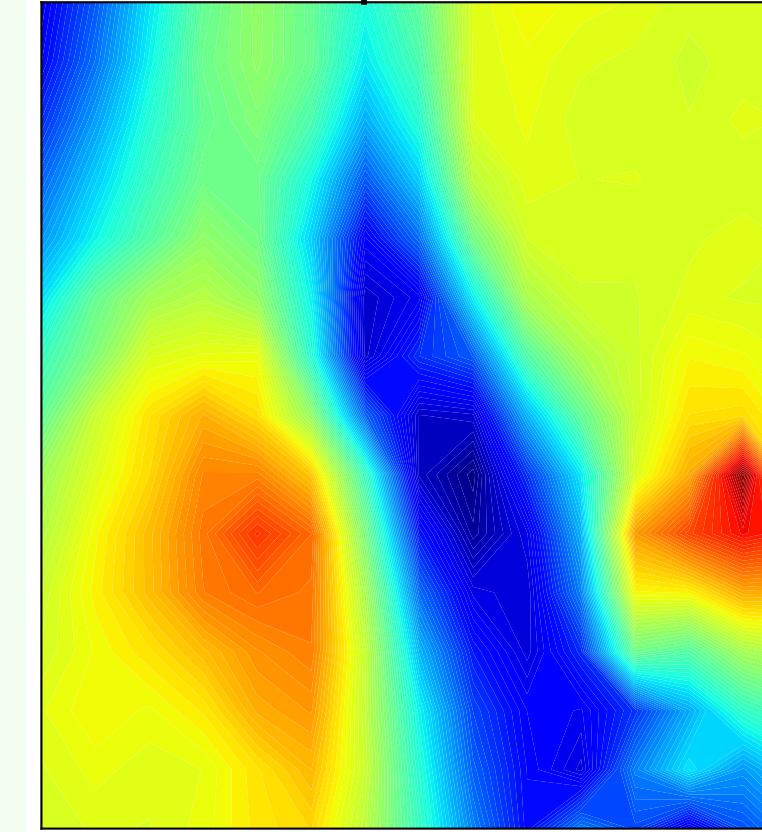
Myopic expected loss



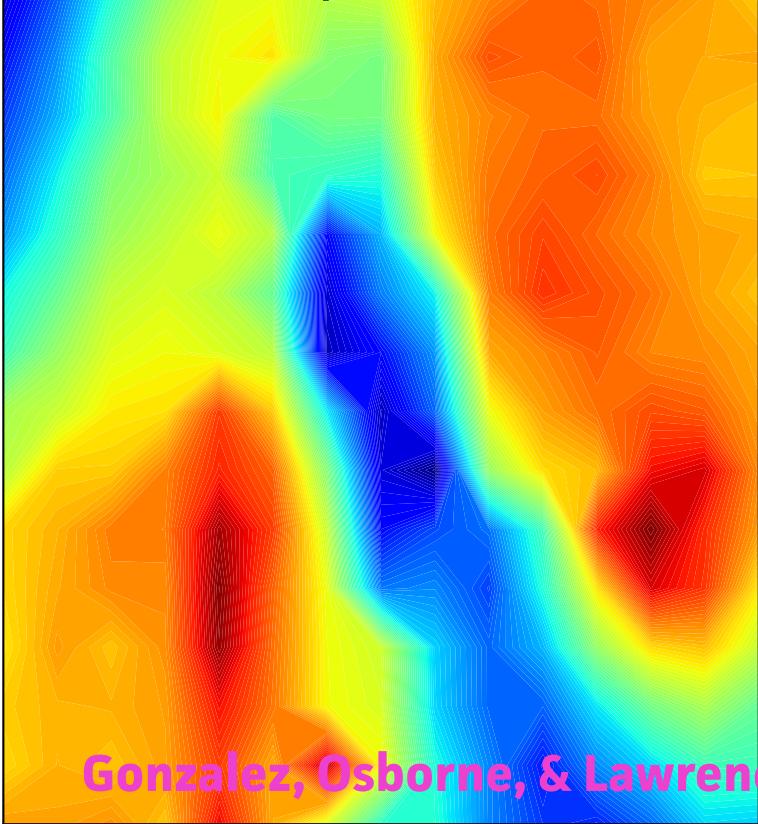
2 steps ahead



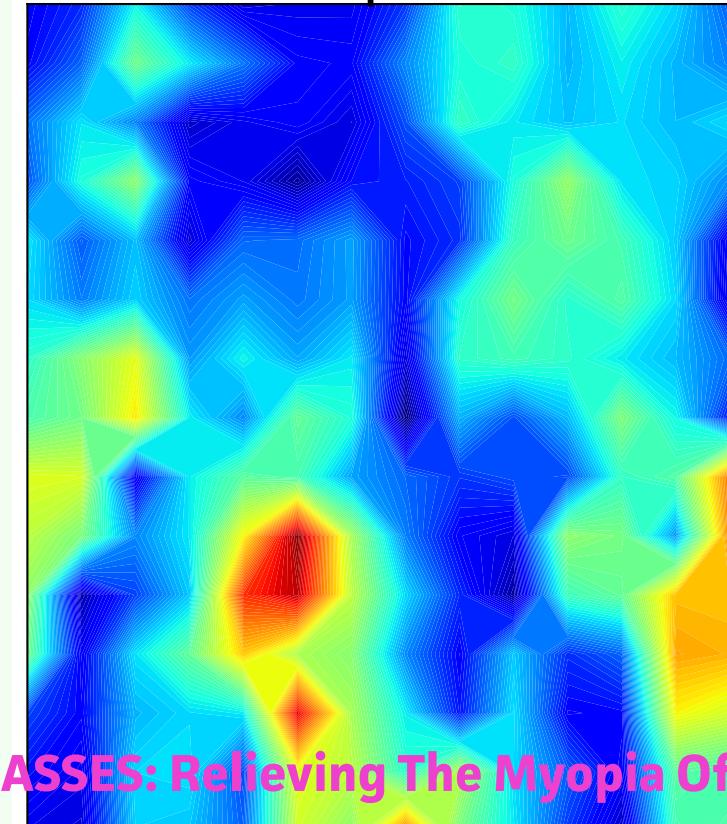
3 steps ahead



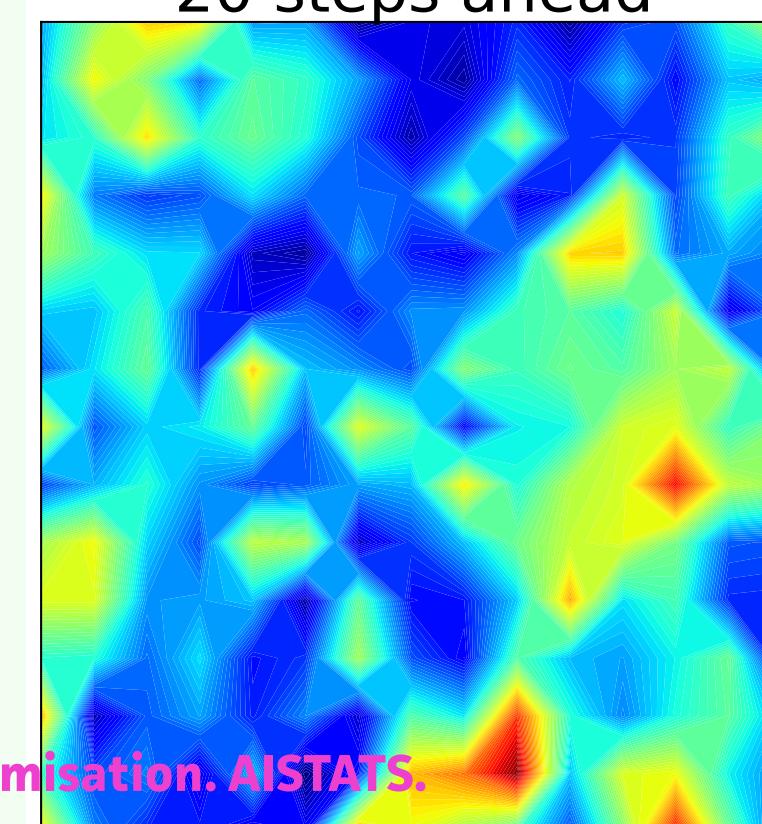
5 steps ahead



10 steps ahead



20 steps ahead



With a myopic strategy, the acquisition function is

$$\begin{aligned}\alpha(\boldsymbol{x}_n \mid \mathcal{D}_n) &= \mathbb{E}(\lambda(\boldsymbol{x}_n, y_n, \mathcal{D}_n)) \\ &= \int \lambda(\boldsymbol{x}_n, y_n, \mathcal{D}_n) p(y_n \mid \mathcal{D}_n) dy_n.\end{aligned}$$

The next evaluation location will be

$$\boldsymbol{x}_n = \arg \min_{\boldsymbol{x}} \alpha(\boldsymbol{x} \mid \mathcal{D}_n).$$

$$\boldsymbol{x}_n = \arg \min_{\boldsymbol{x}} \alpha(\boldsymbol{x} \mid \mathcal{D}_n).$$

We have succeeded
in turning optimisation
into optimisation.

The acquisition function:

is less expensive than the objective;

gives us gradients and Hessians;
and

need not be optimised exactly.

Expected improvement

is a myopic approximation to the value loss:

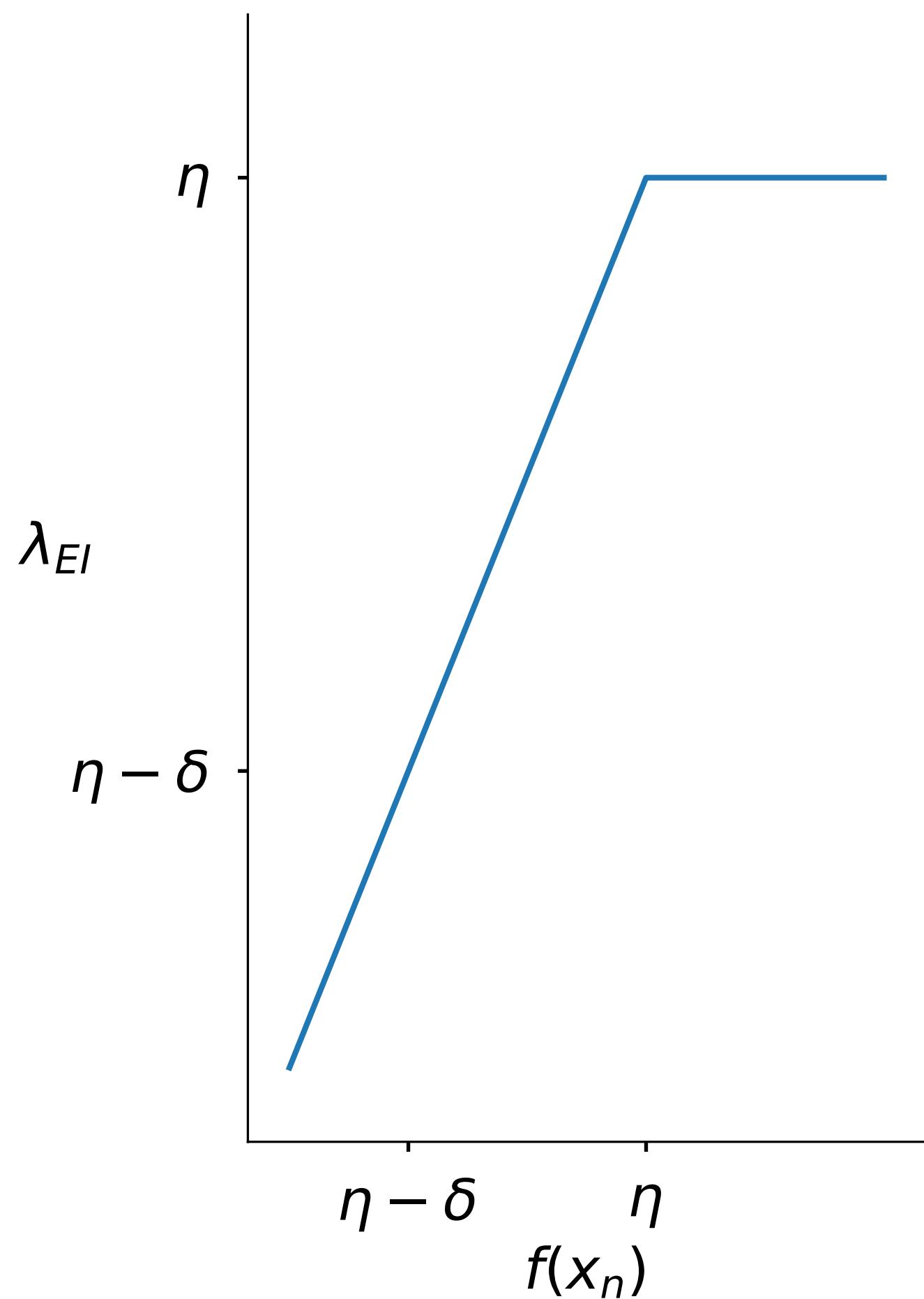
$$\begin{aligned}\lambda_{\text{VL}}(\mathbf{x}_N, f(\mathbf{x}_N), \mathcal{D}_N) \\ \simeq \lambda_{\text{EI}}(\mathcal{D}_{n+1}) \\ := \min_{i \in \{0, \dots, n\}} f(\mathbf{x}_i).\end{aligned}$$

Defining the lowest function value available at the n th step as

$$\eta := \min_{i \in \{0, \dots, n-1\}} f(\mathbf{x}_i),$$

we can simply rewrite the loss as

$$\lambda_{\text{EI}}(\mathcal{D}_{n+1}) = \min\{\eta, f(\mathbf{x}_n)\}.$$



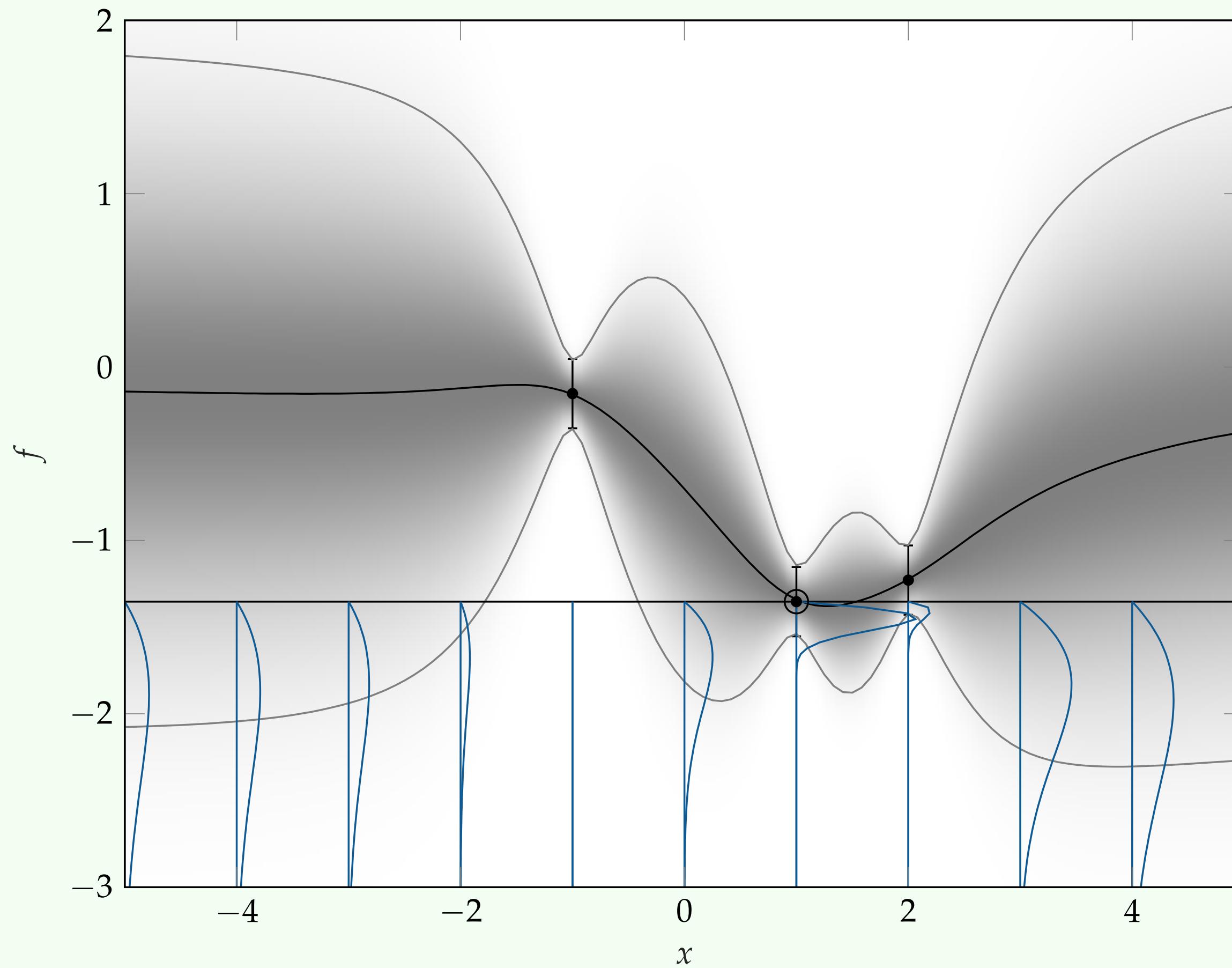
If we have a Gaussian posterior for the next evaluation,

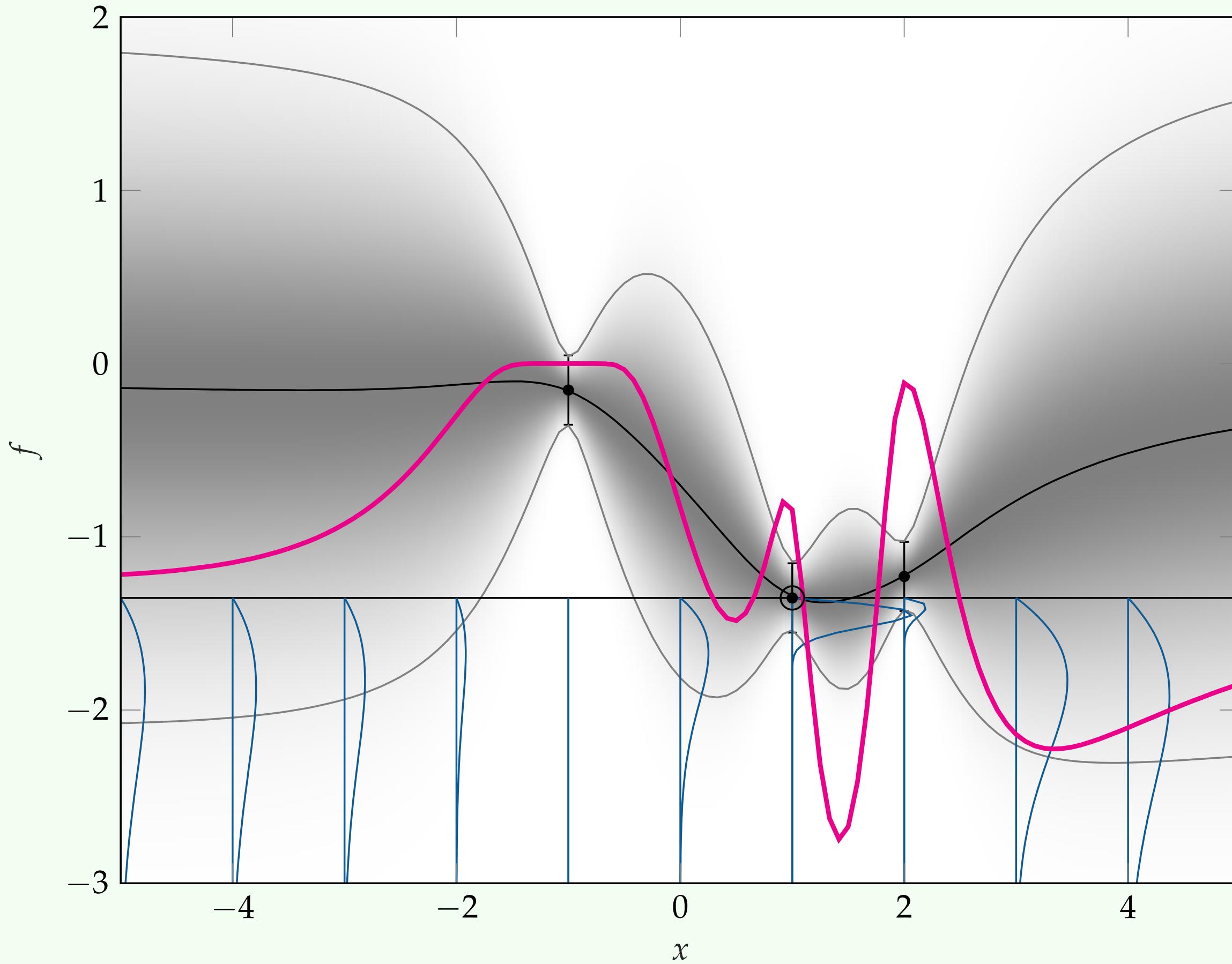
$$p(f(\mathbf{x}_n) \mid \mathcal{D}_n) := \mathcal{N}(f(\mathbf{x}_n); m(\mathbf{x}_n), V(\mathbf{x}_n)),$$

the **expected improvement acquisition function** is

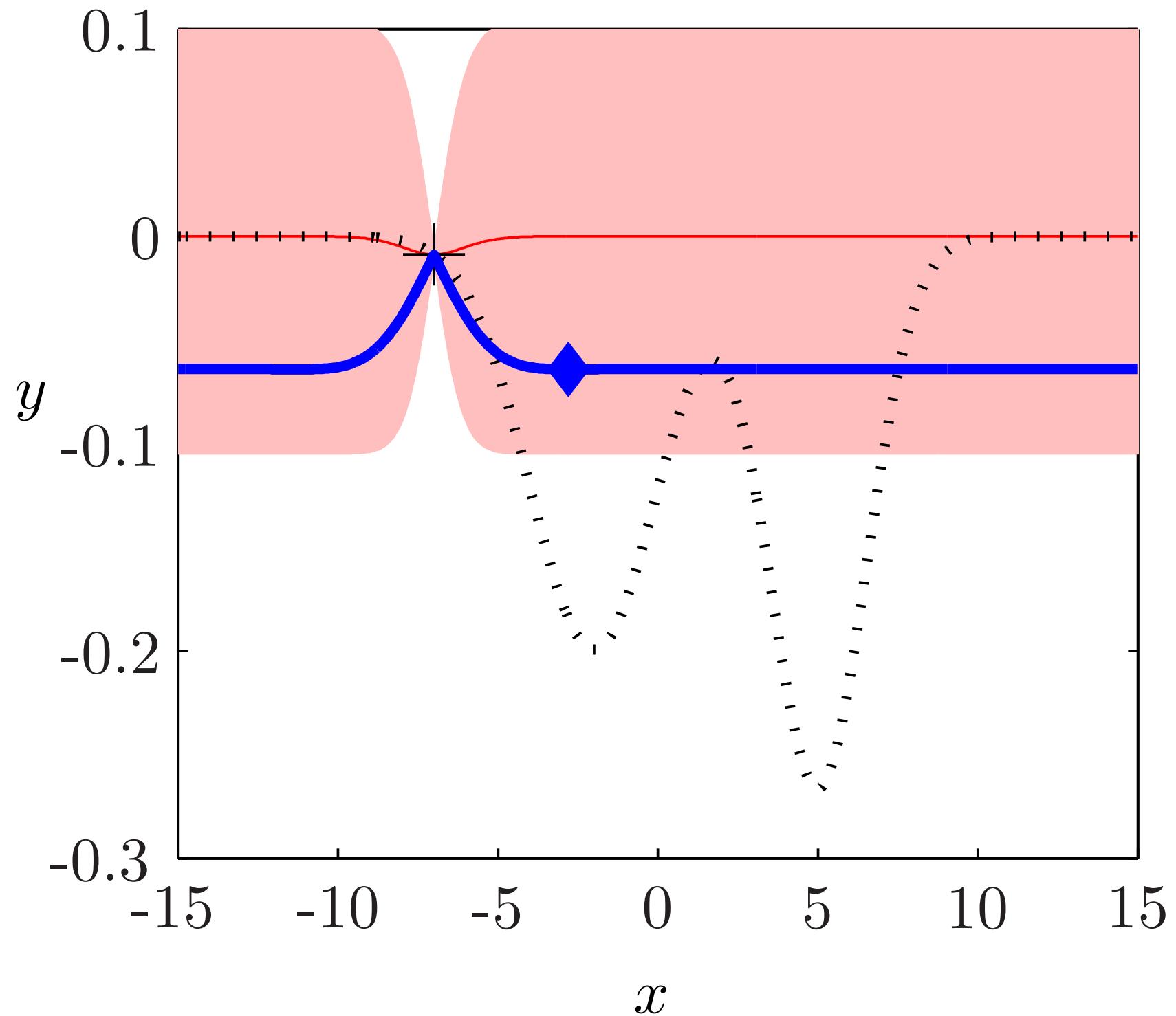
$$\begin{aligned}\alpha_{\text{EI}}(\mathbf{x}_n) &:= \mathbb{E}(\lambda_{\text{EI}})(\mathbf{x}_n) - \eta \\ &= \int_{-\infty}^{\eta} (f(\mathbf{x}_n) - \eta) p(f(\mathbf{x}_n) \mid \mathcal{D}_n) df(\mathbf{x}_n) \\ &= -V(\mathbf{x}_n) \mathcal{N}(\eta; m(\mathbf{x}_n), V(\mathbf{x}_n)) \\ &\quad + (m(\mathbf{x}_n) - \eta) \Phi(\eta; m(\mathbf{x}_n), V(\mathbf{x}_n)).\end{aligned}$$

$$\alpha_{\text{EI}}(\boldsymbol{x}_n) = \int_{-\infty}^{\eta} (f(\boldsymbol{x}_n) - \eta) p(f(\boldsymbol{x}_n) \mid \mathcal{D}_n) \, df(\boldsymbol{x}_n)$$



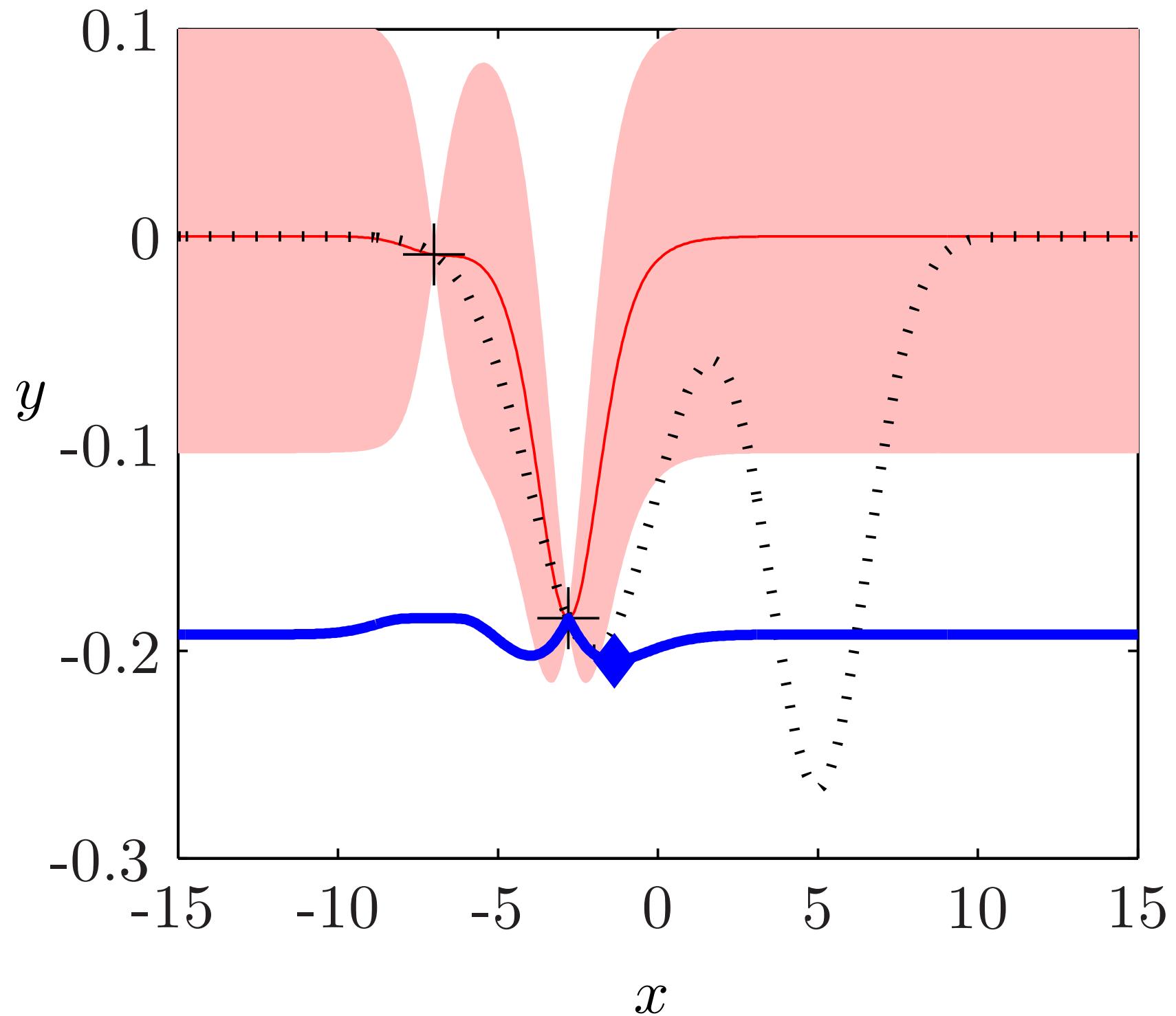


Function Evaluation 1



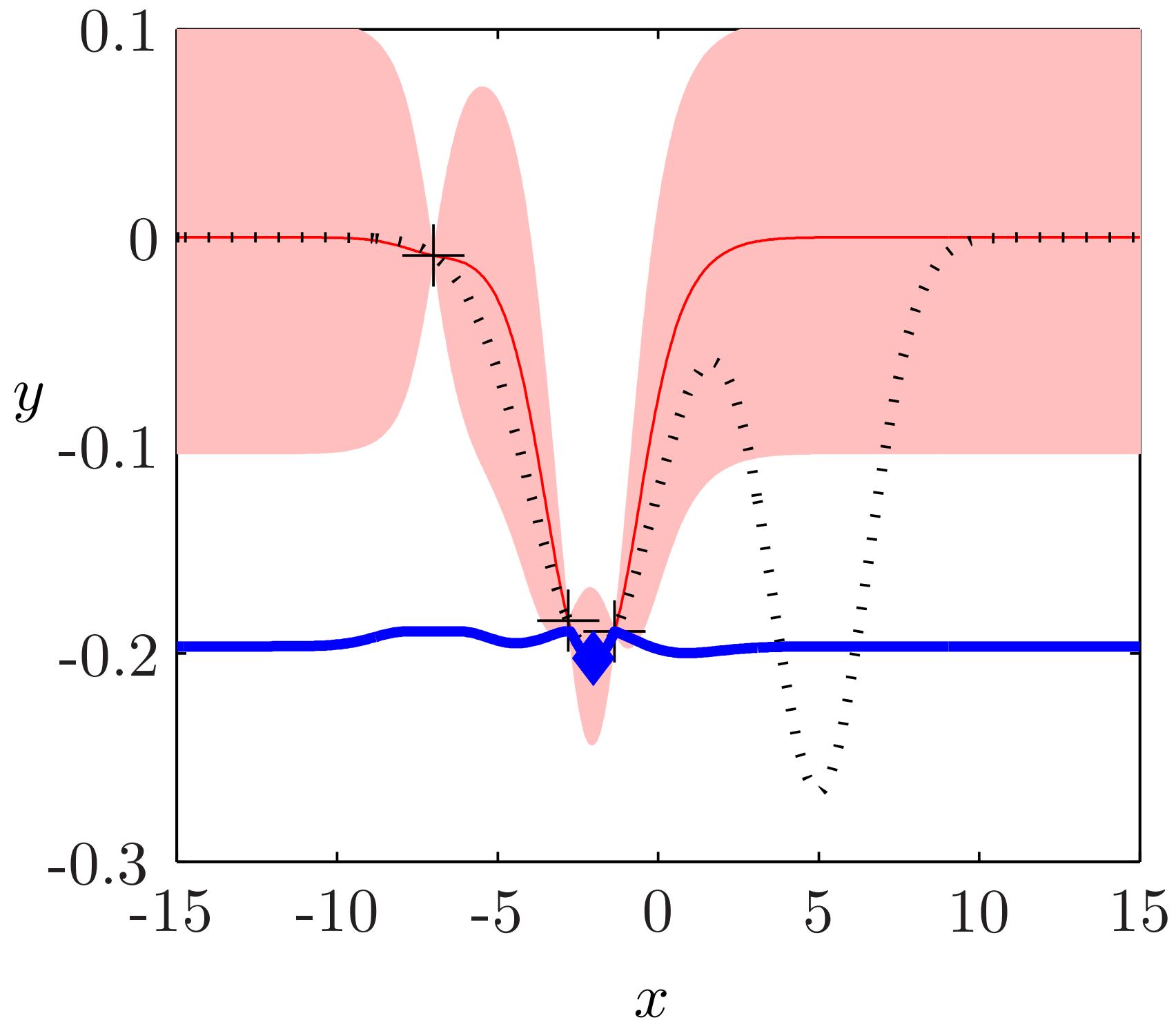
- Objective function
- + Observation
- Mean
- ± 1SD
- Expected loss
- ◆ Chosen position of next observation

Function Evaluation 2



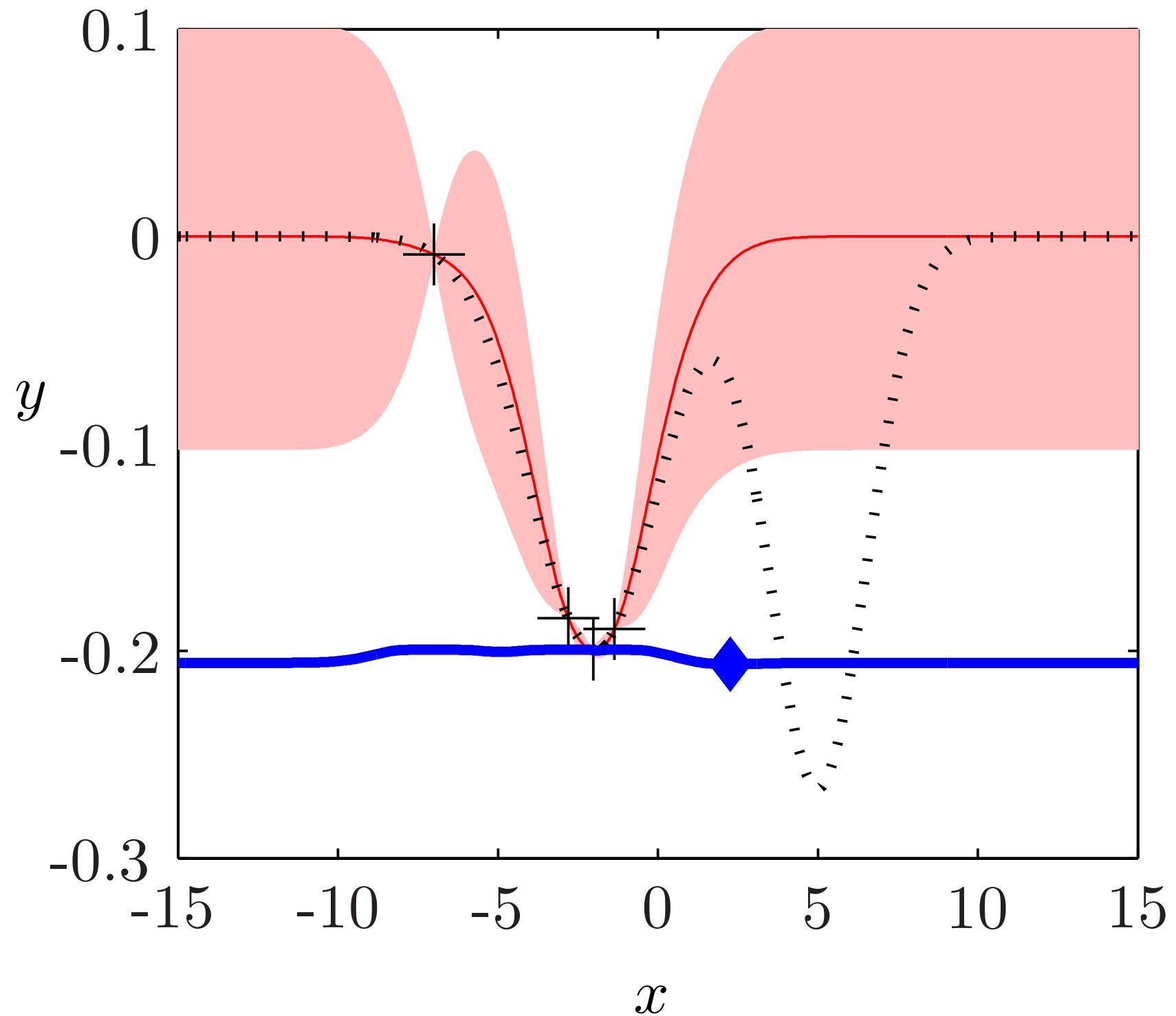
- Objective function
- + Observation
- Mean
- ± 1SD
- Expected loss
- ◆ Chosen position of next observation

Function Evaluation 3



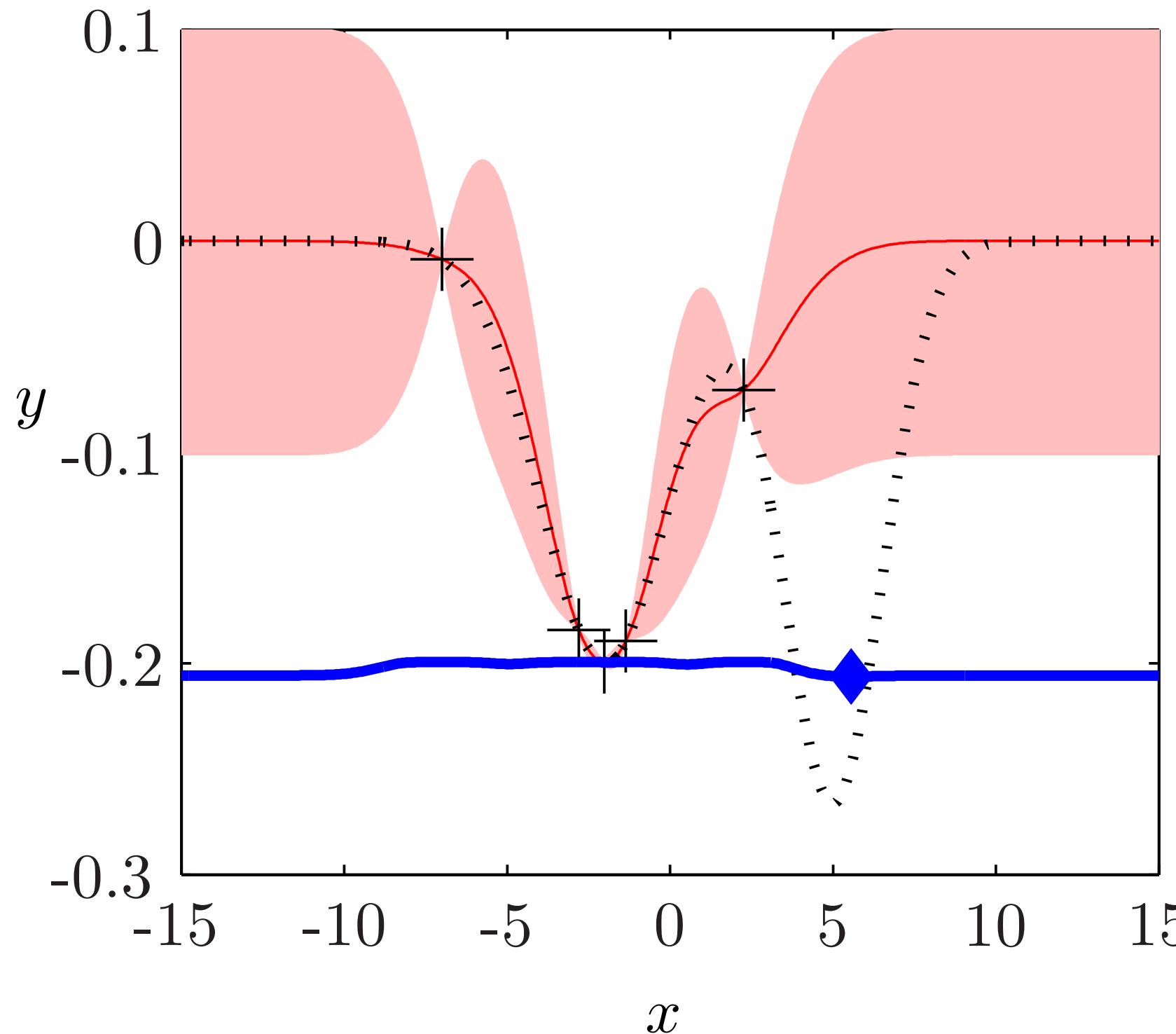
- Objective function
- + Observation
- Mean
- ± 1SD
- Expected loss
- ◆ Chosen position of next observation

Function Evaluation 4



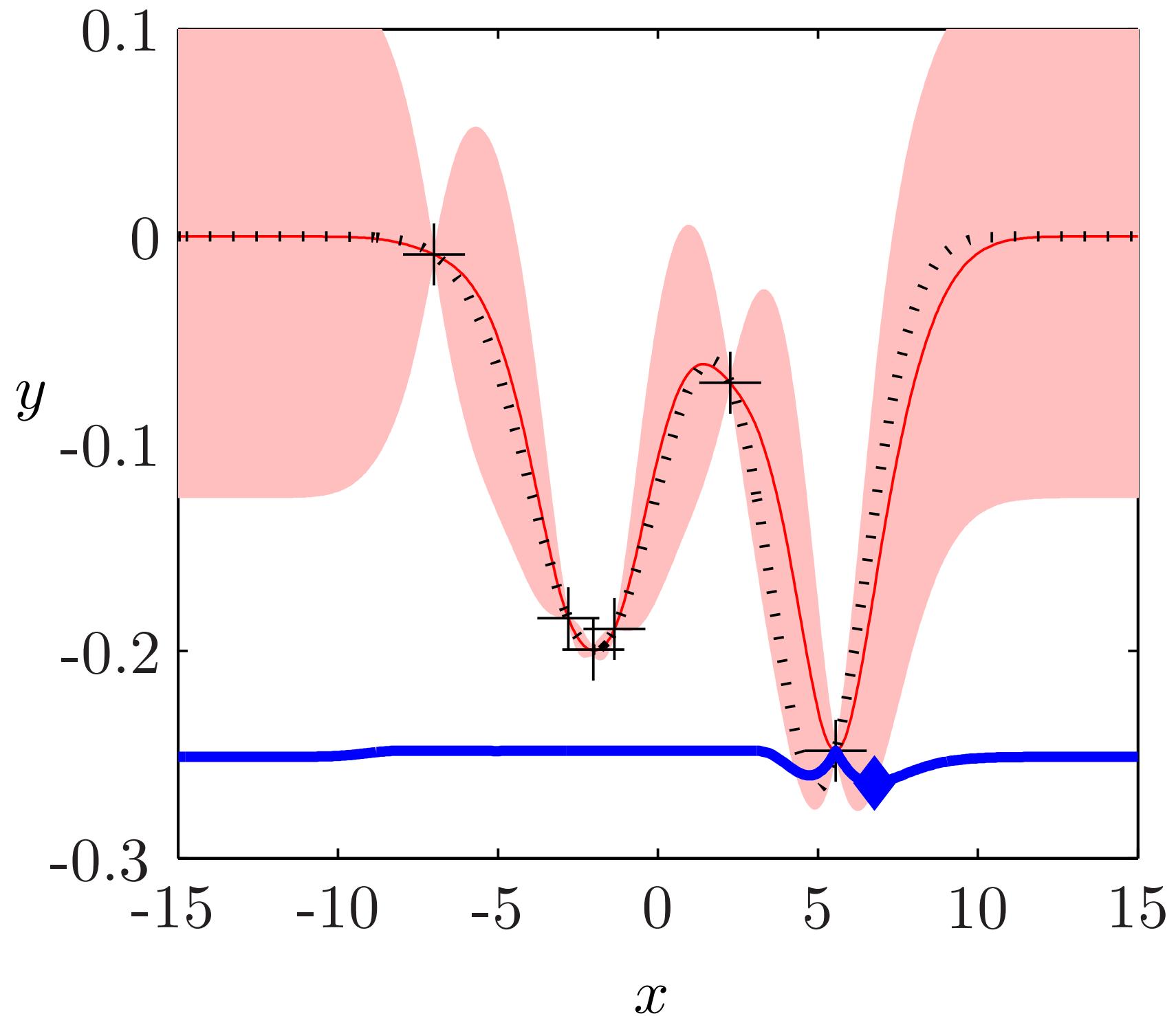
- Objective function
- + Observation
- Mean
- ± 1SD
- Expected loss
- ◆ Chosen position of next observation

Function Evaluation 5



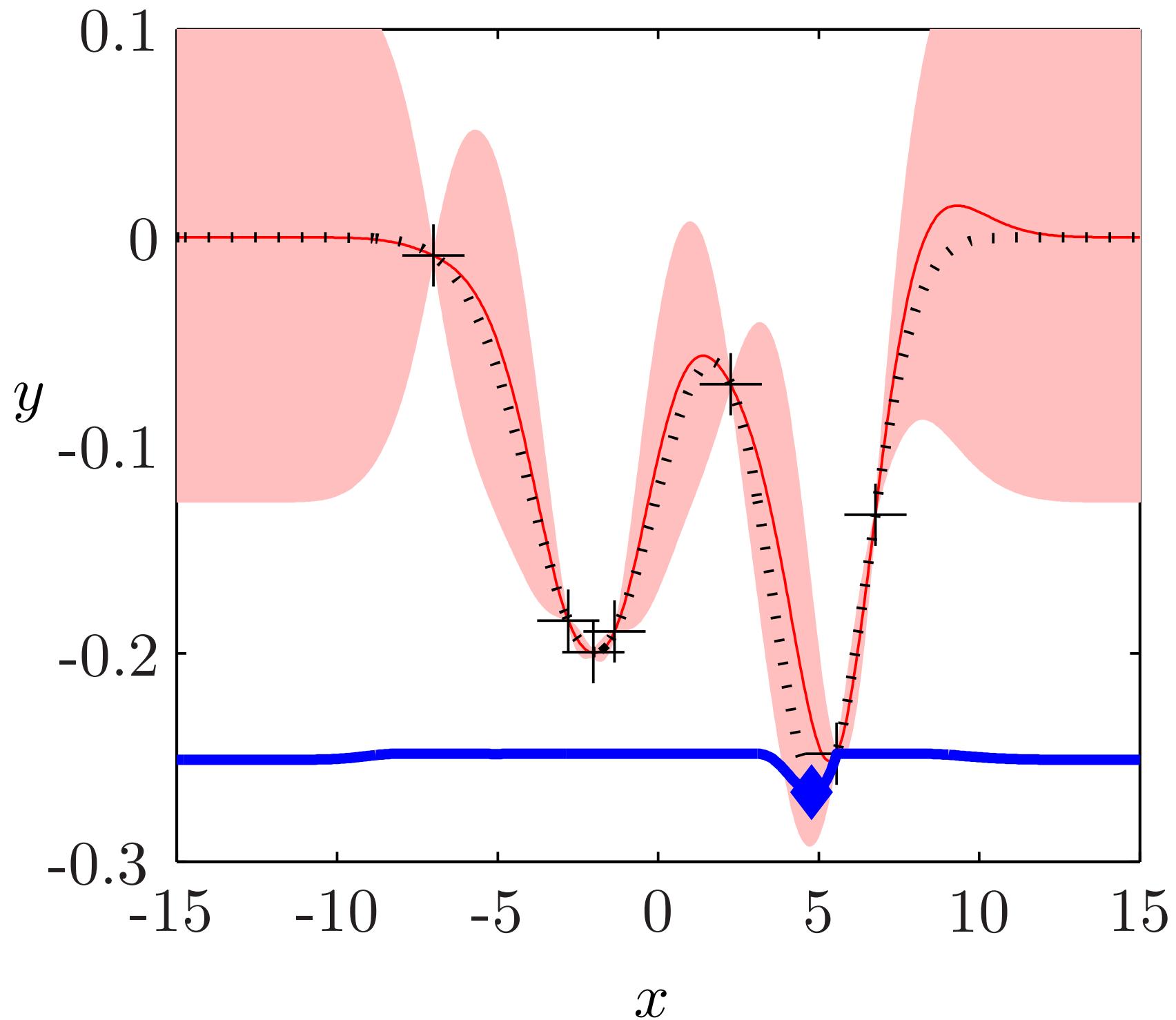
- Objective function
- + Observation
- Mean
- ± 1SD
- Expected loss
- ◆ Chosen position of next observation

Function Evaluation 6



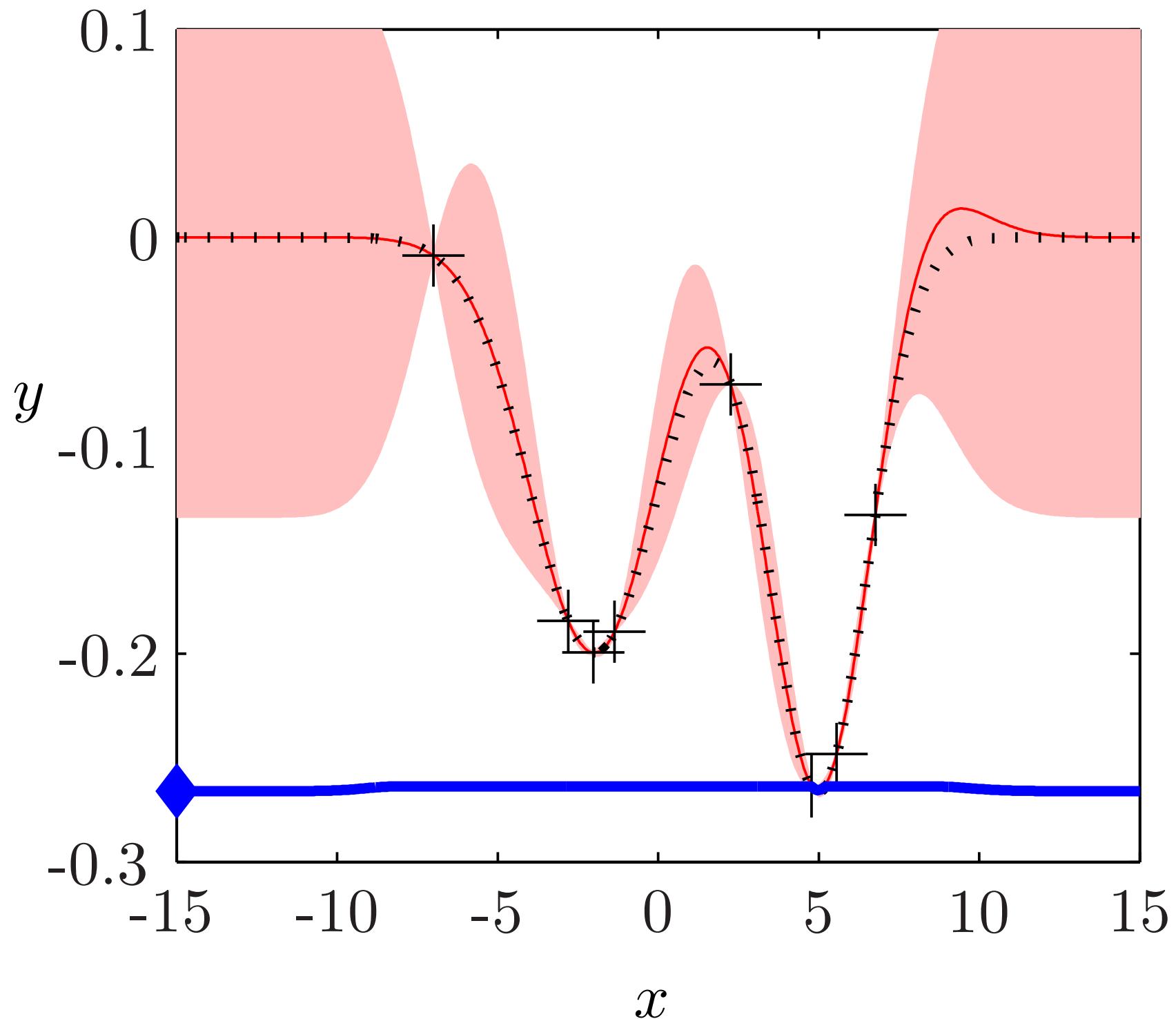
- Objective function
- + Observation
- Mean
- ± 1SD
- Expected loss
- ◆ Chosen position of next observation

Function Evaluation 7



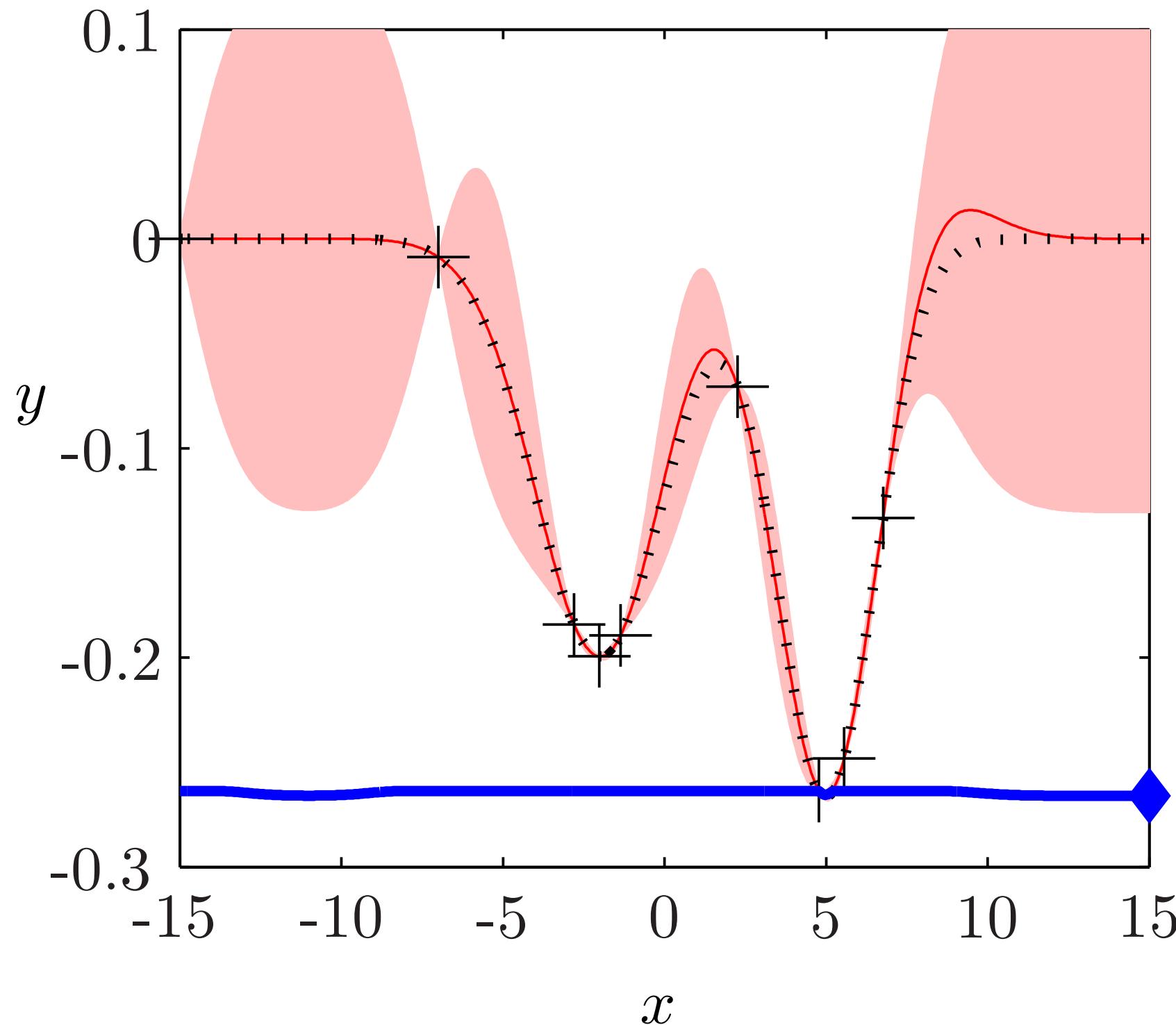
- Objective function
- + Observation
- Mean
- ± 1SD
- Expected loss
- ◆ Chosen position of next observation

Function Evaluation 8



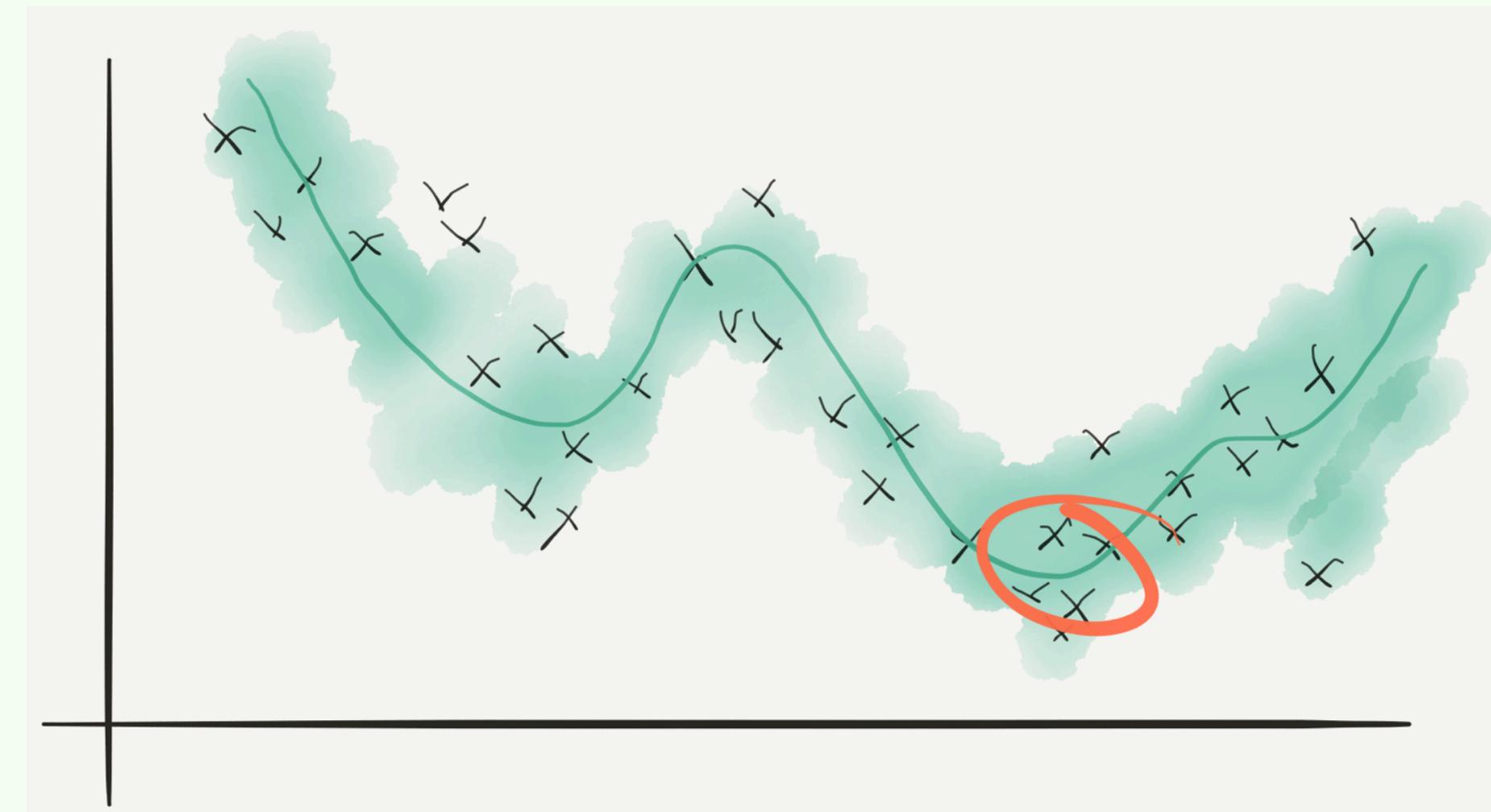
- Objective function
- + Observation
- Mean
- ± 1SD
- Expected loss
- ◆ Chosen position of next observation

Function Evaluation 9



- Objective function
- + Observation
- Mean
- ± 1SD
- Expected loss
- ◆ Chosen position of next observation

If our evaluations are **noisy**, the best evaluation (η) is also probably the most **noise-corrupted**.



Probability of improvement

defines (for \mathbb{I} the indicator function) the myopic loss

$$\lambda_{n,\text{PI}}(\mathcal{D}_{n+1}) := \mathbb{I}(f(\boldsymbol{x}_n) \geq \eta).$$

The probability of improvement acquisition function is hence

$$\alpha_{n,\text{PI}}(\boldsymbol{x}_n) := \mathbb{E}(\lambda_{n,\text{PI}}(\mathcal{D}_{n+1})) = P(f(\boldsymbol{x}_n) \geq \eta \mid \mathcal{D}_n).$$

Probability of improvement

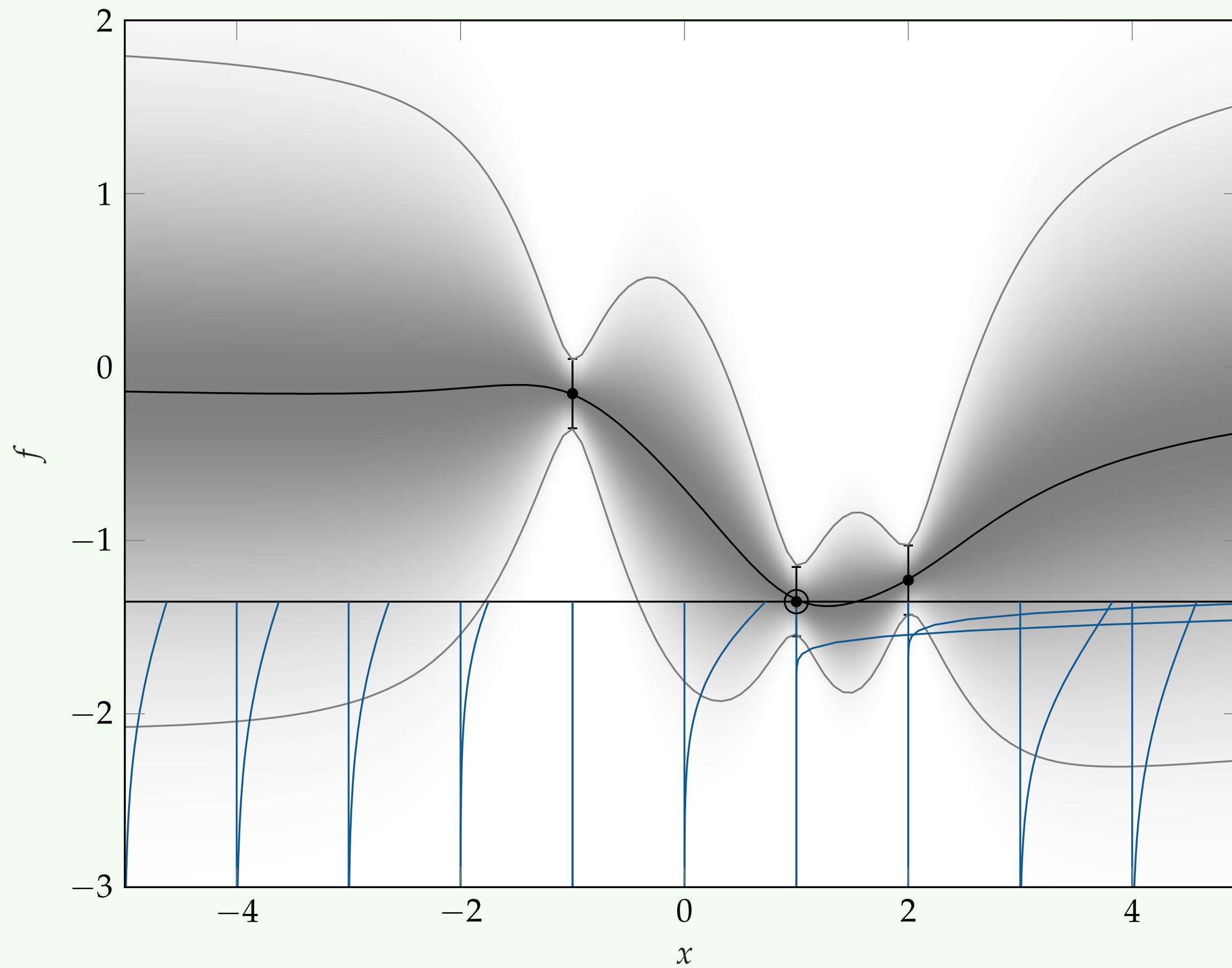
defines a myopic loss (for \mathbb{I} the indicator function)

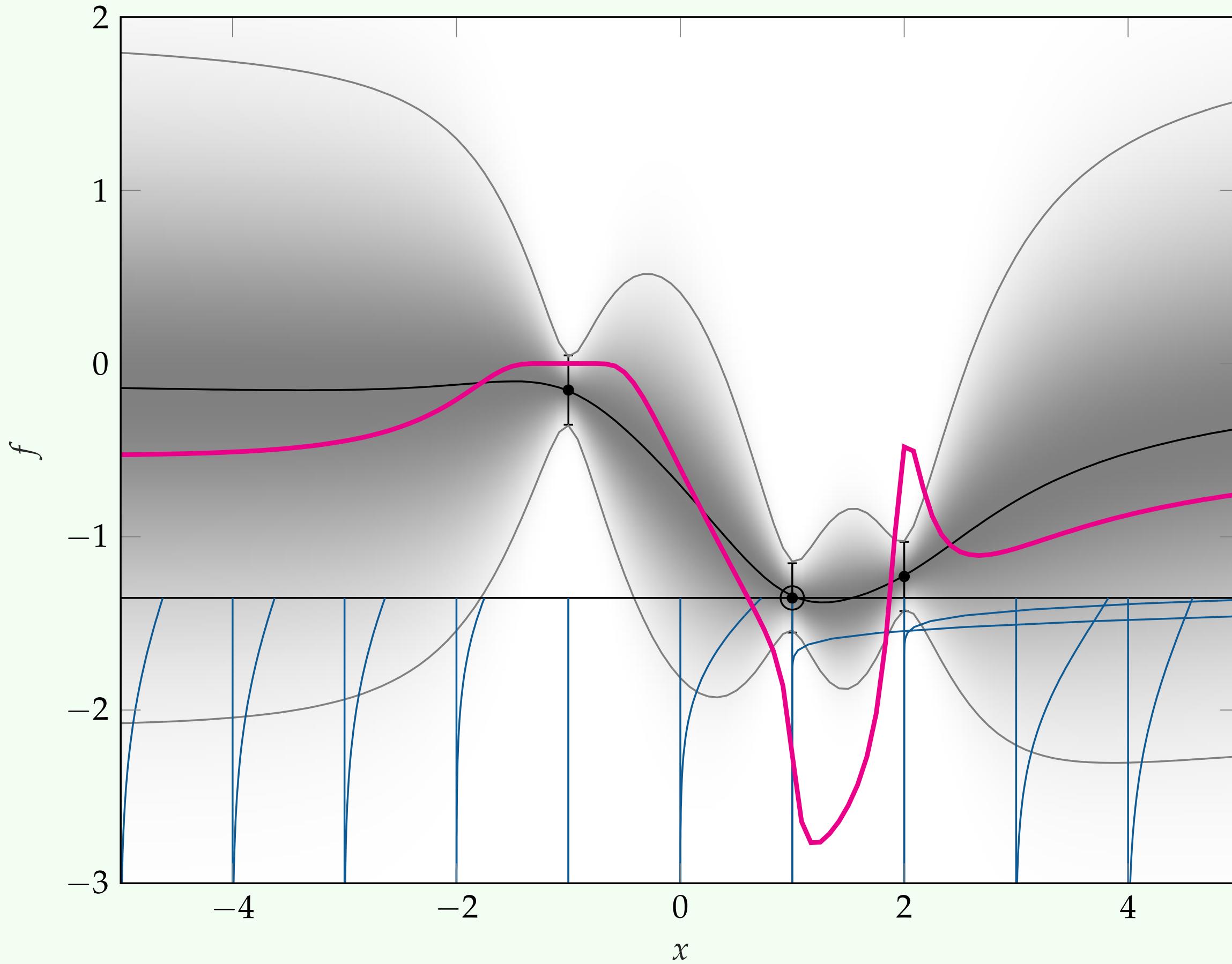
$$\lambda_{n,\text{PI}}(\mathcal{D}_{n+1}) := \mathbb{I}(f(\boldsymbol{x}_n) \geq \eta).$$

The probability of improvement acquisition function is hence

$$\alpha_{n,\text{PI}}(\boldsymbol{x}_n) := \mathbb{E}(\lambda_{n,\text{PI}}(\mathcal{D}_{n+1})) = P(f(\boldsymbol{x}_n) \geq \eta \mid \mathcal{D}_n).$$

PI values incremental improvement every step.





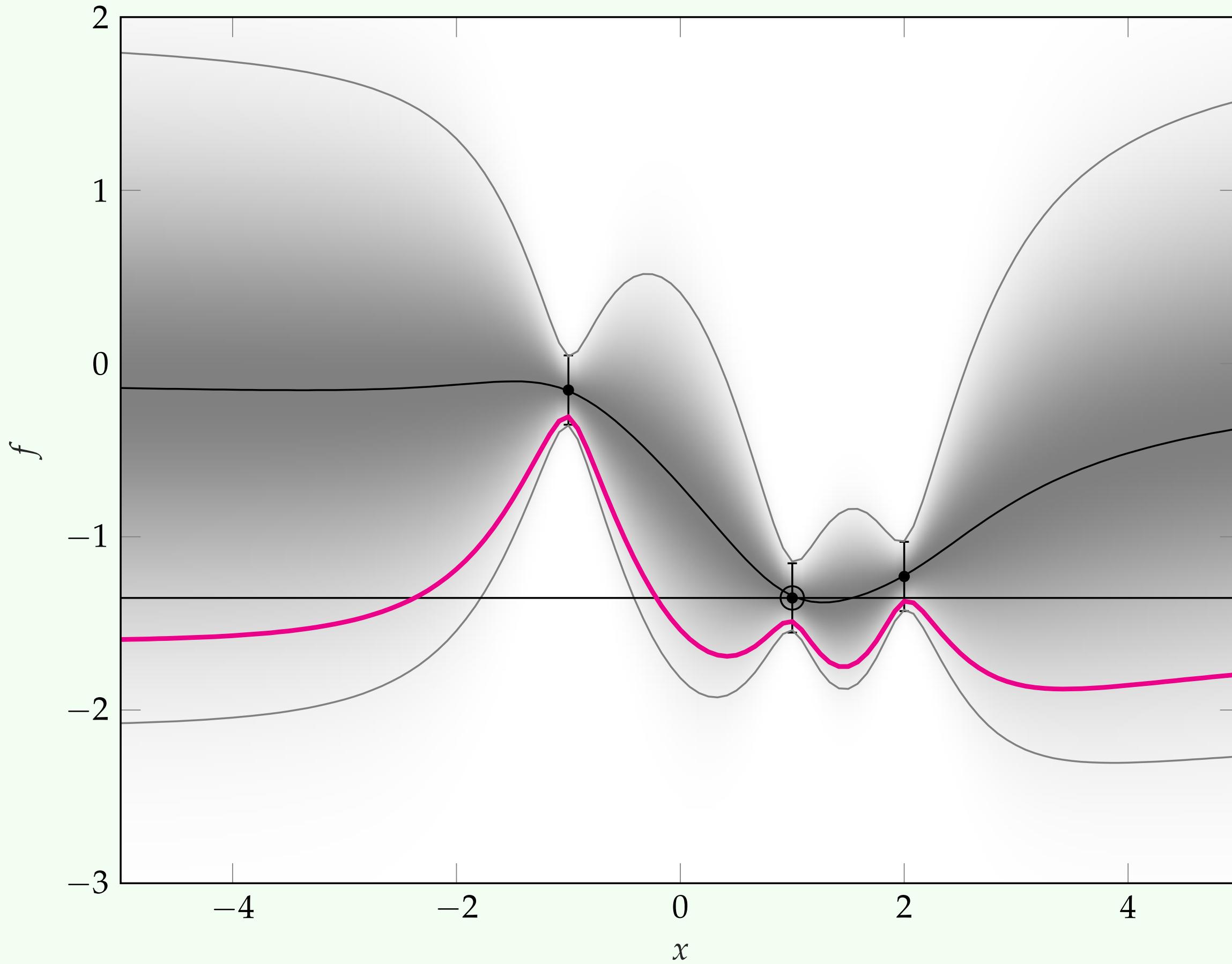
Upper confidence bound

is the myopic acquisition function

$$\alpha_{\text{UCB}}(\boldsymbol{x}_n) := m(\boldsymbol{x}_n) - \beta_n V(\boldsymbol{x}_n)^{\frac{1}{2}}.$$

given a surrogate with mean $m(\boldsymbol{x}_n)$ and variance $V(\boldsymbol{x}_n)$.

It is difficult to reconcile UCB with a defensible loss function.

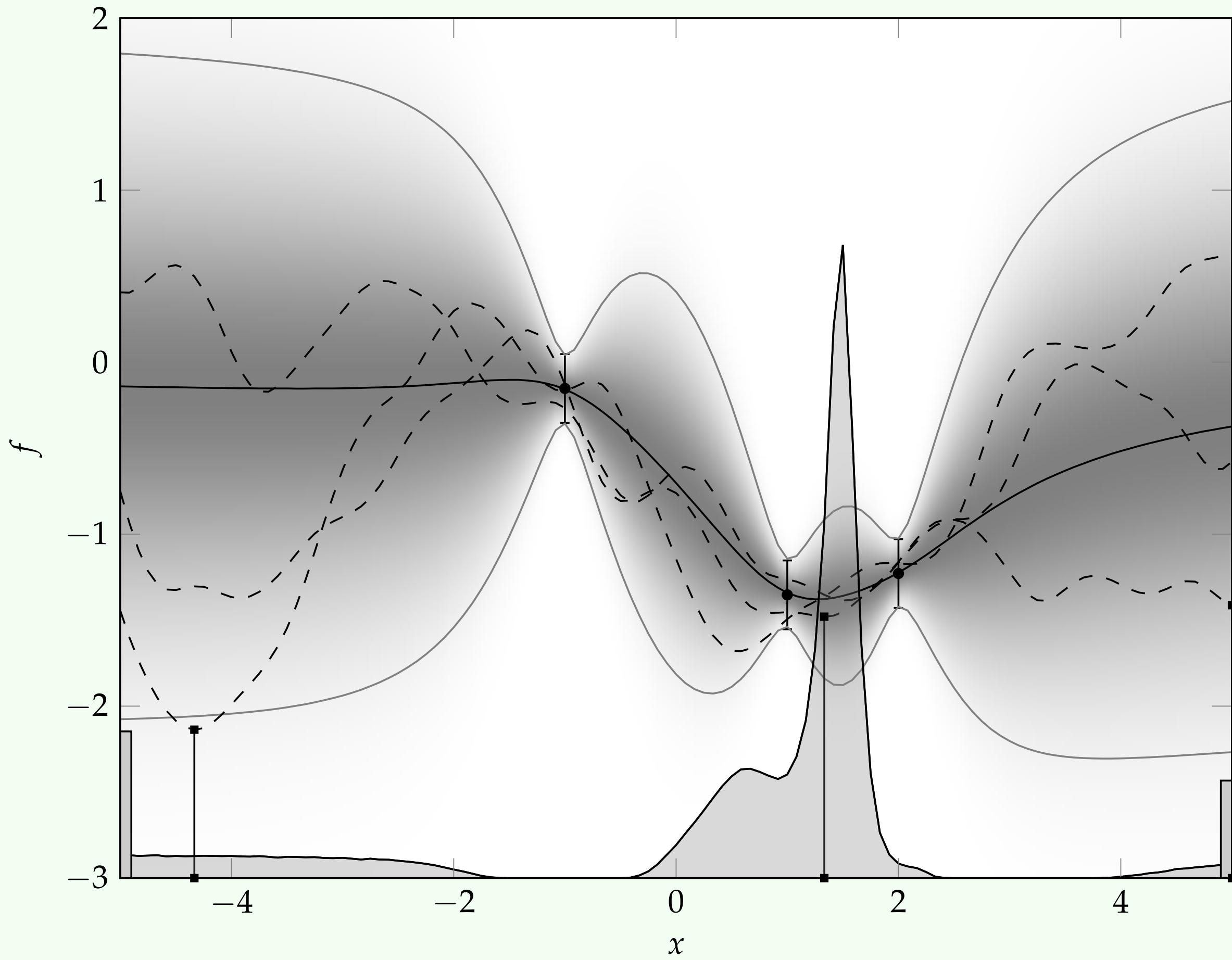


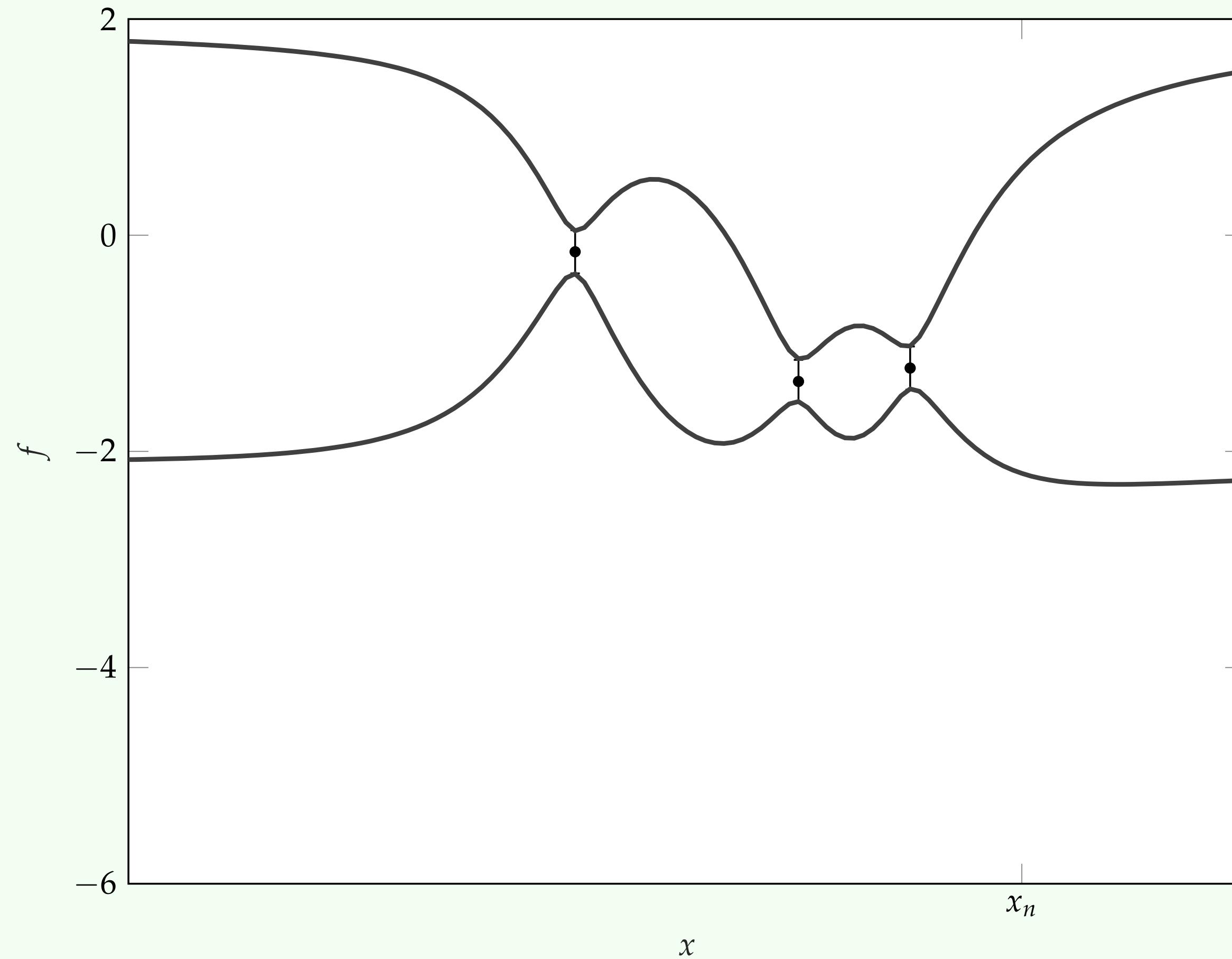
Information-theoretic methods

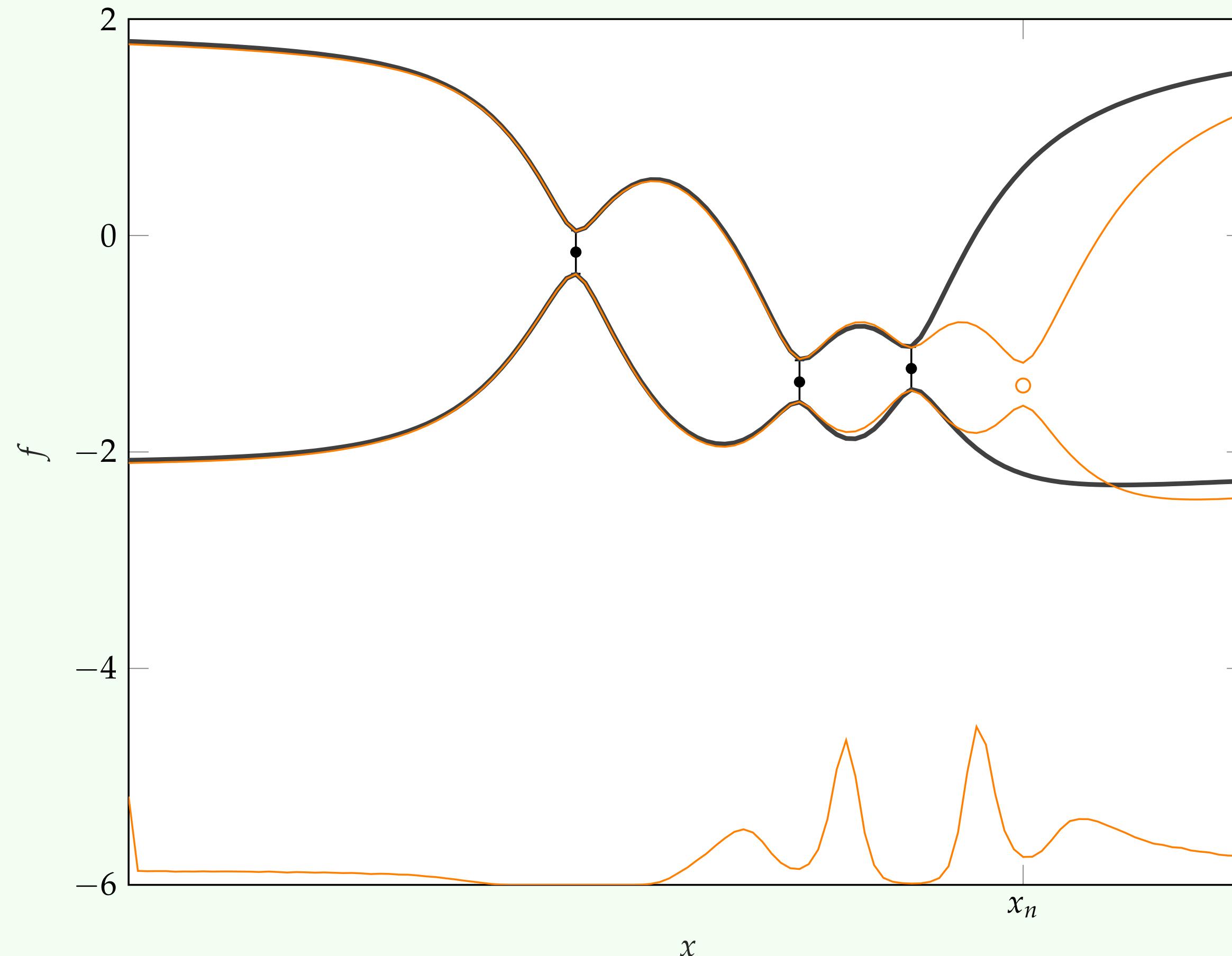
give alternative myopic implementations of value-information and location-information losses:

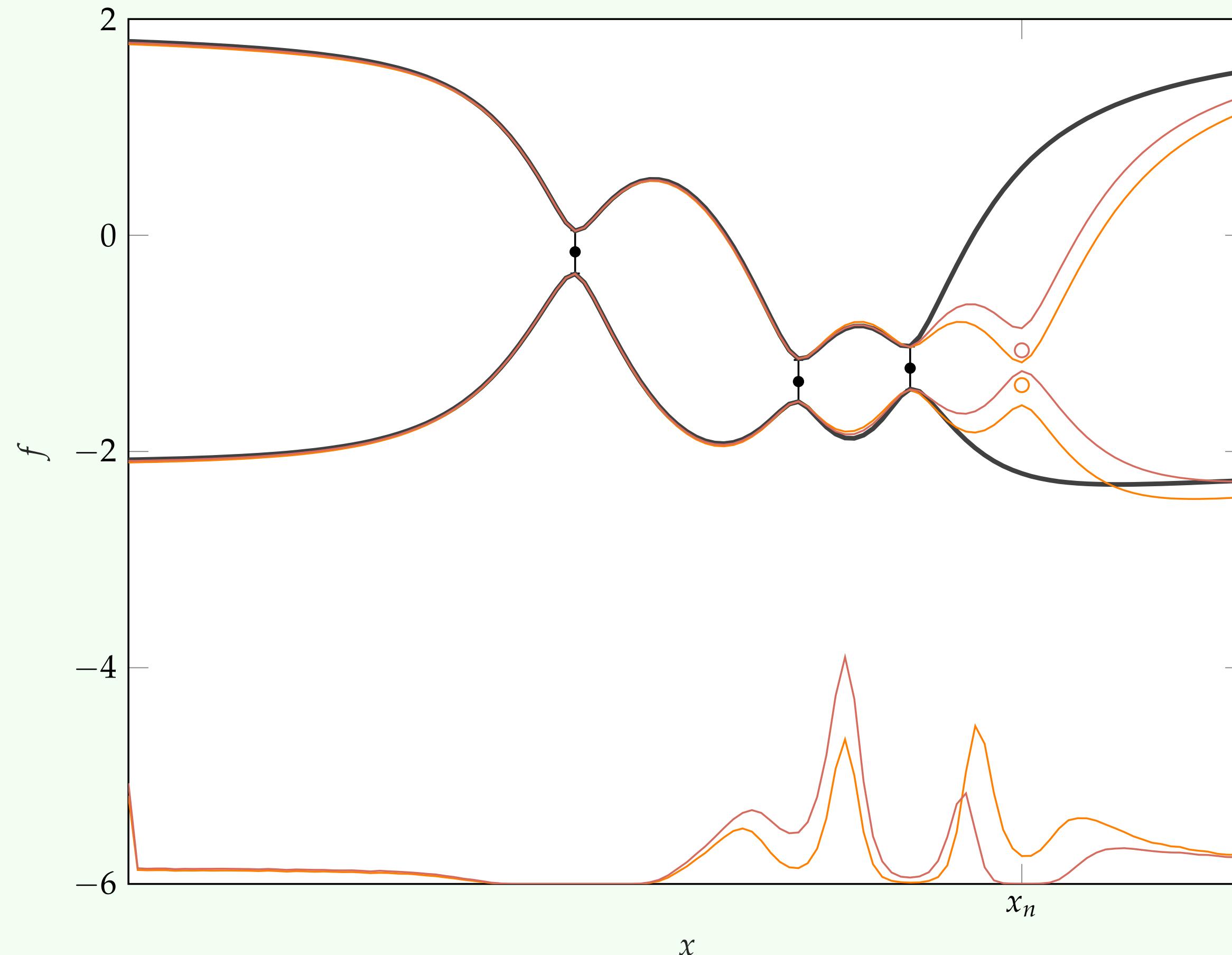
$$\begin{aligned}\alpha_{\text{LIL}} &:= \mathbb{E}_{y_n} \mathbb{H}(\mathbf{x}_* \mid y_n, \mathbf{x}_n, \mathcal{D}_n) \quad \text{and} \\ \alpha_{\text{VIL}} &:= \mathbb{E}_{y_n} \mathbb{H}(y_* \mid y_n, \mathbf{x}_n, \mathcal{D}_n).\end{aligned}$$

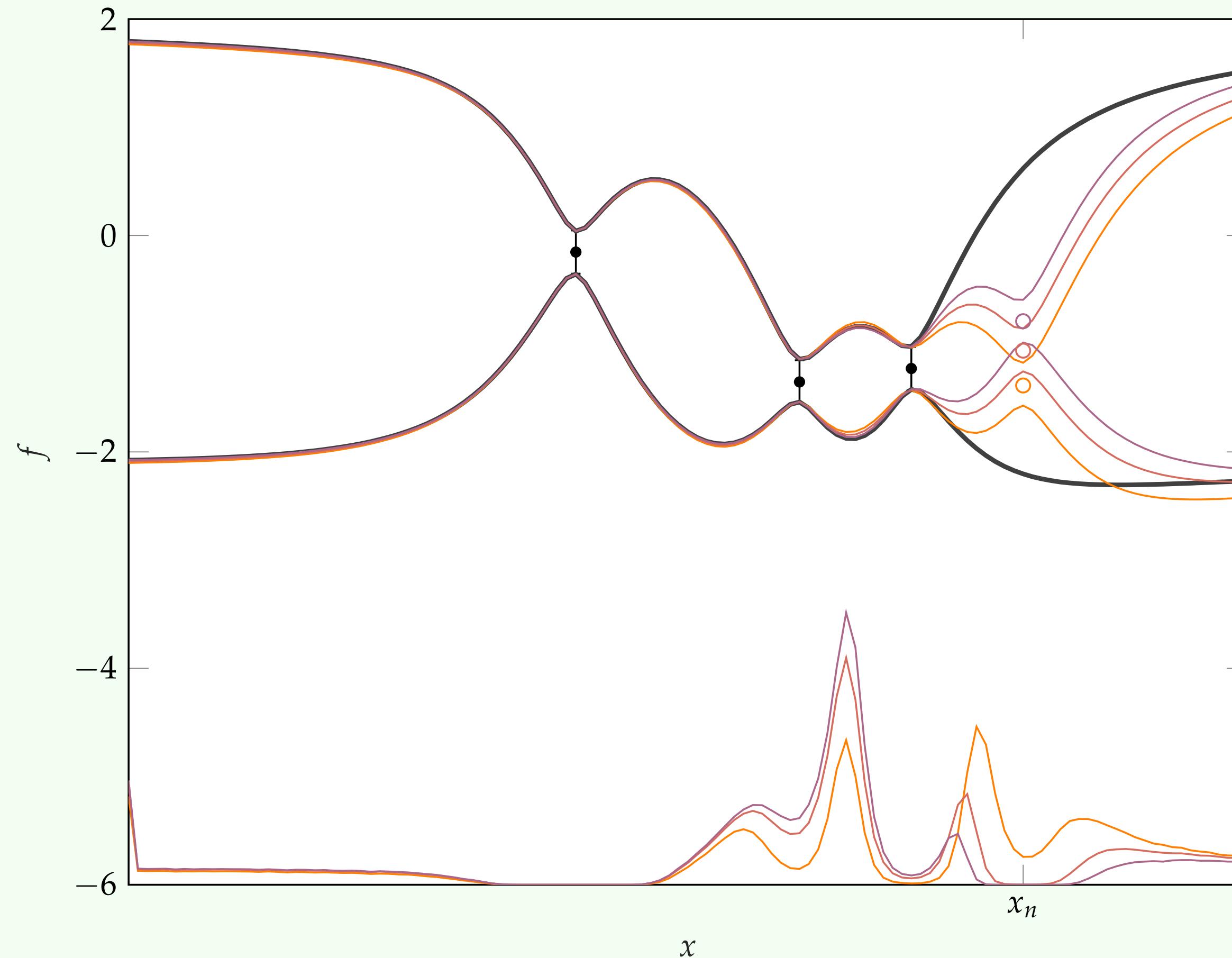
These methods tend to be more exploratory, helping performance.

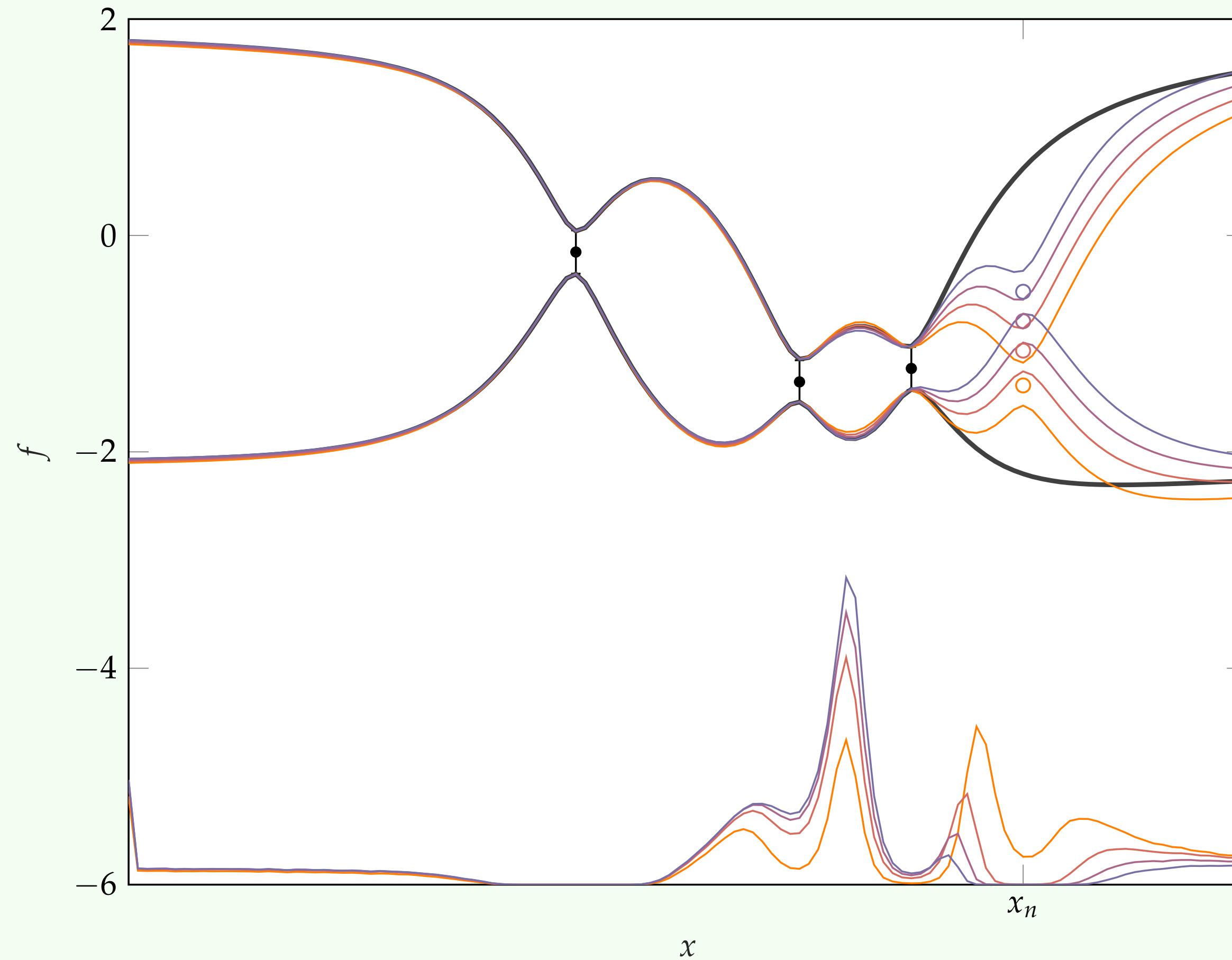


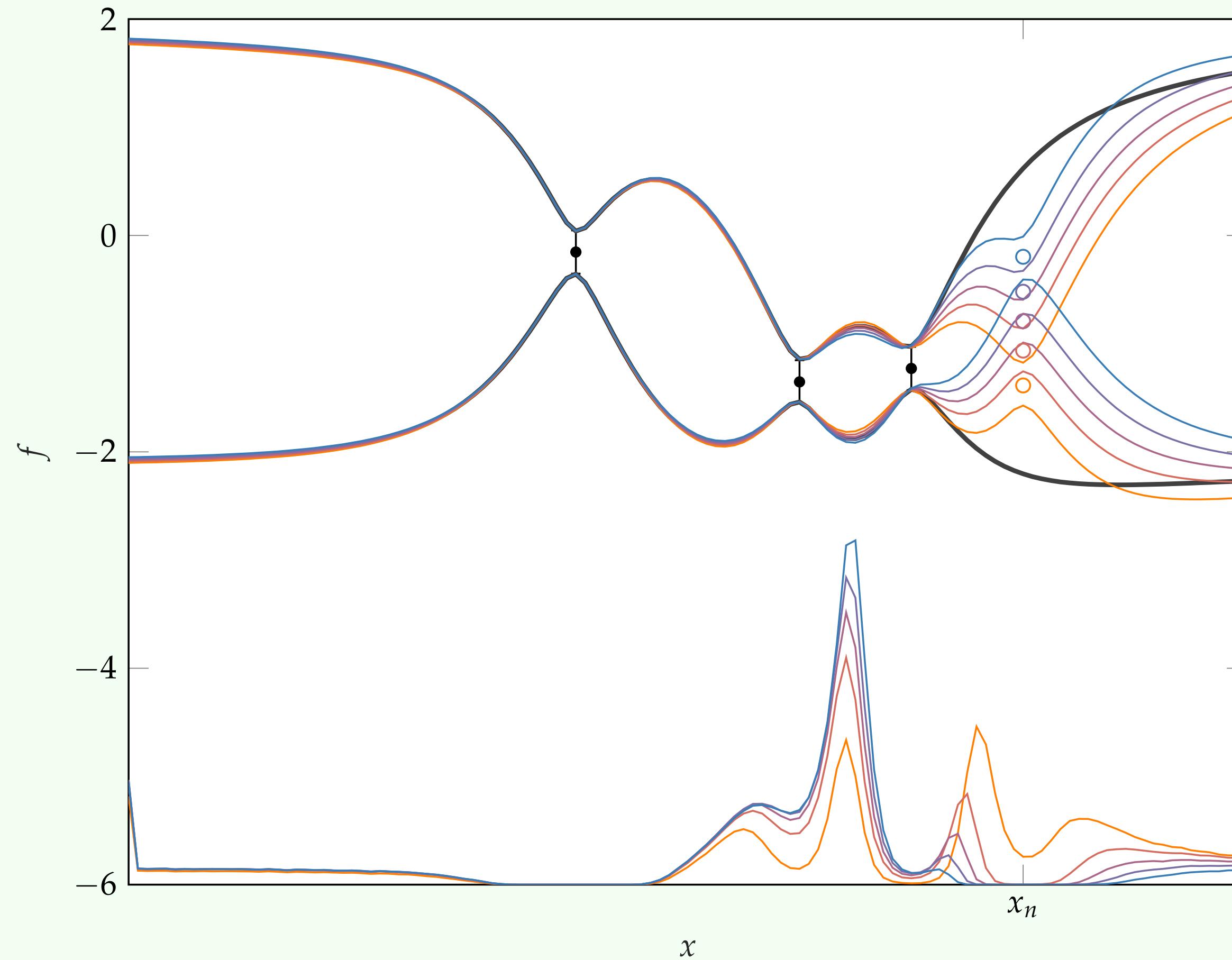


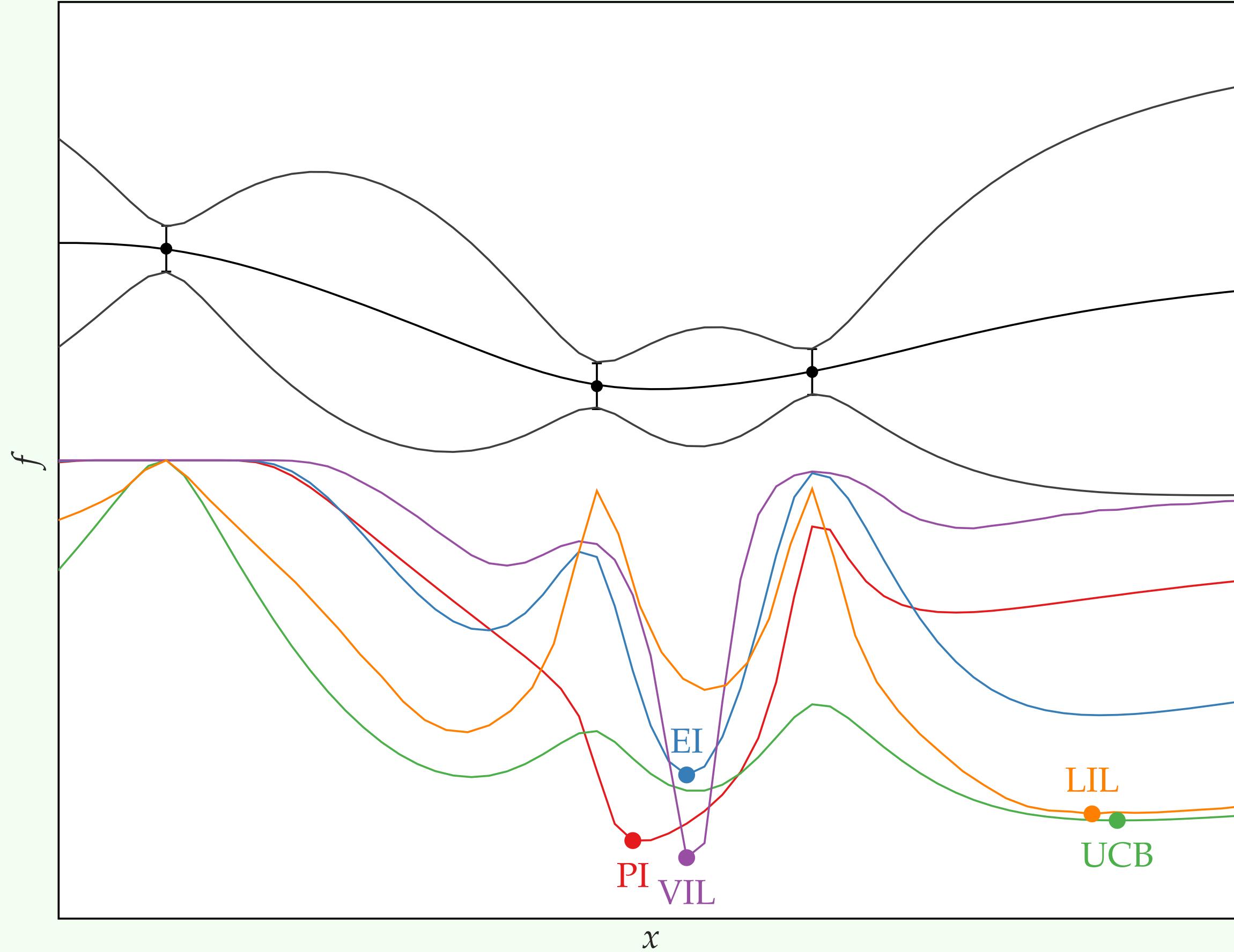










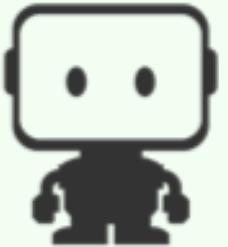


Bayesian optimisation of hyperparameters is used in AutoML.



Whetlab

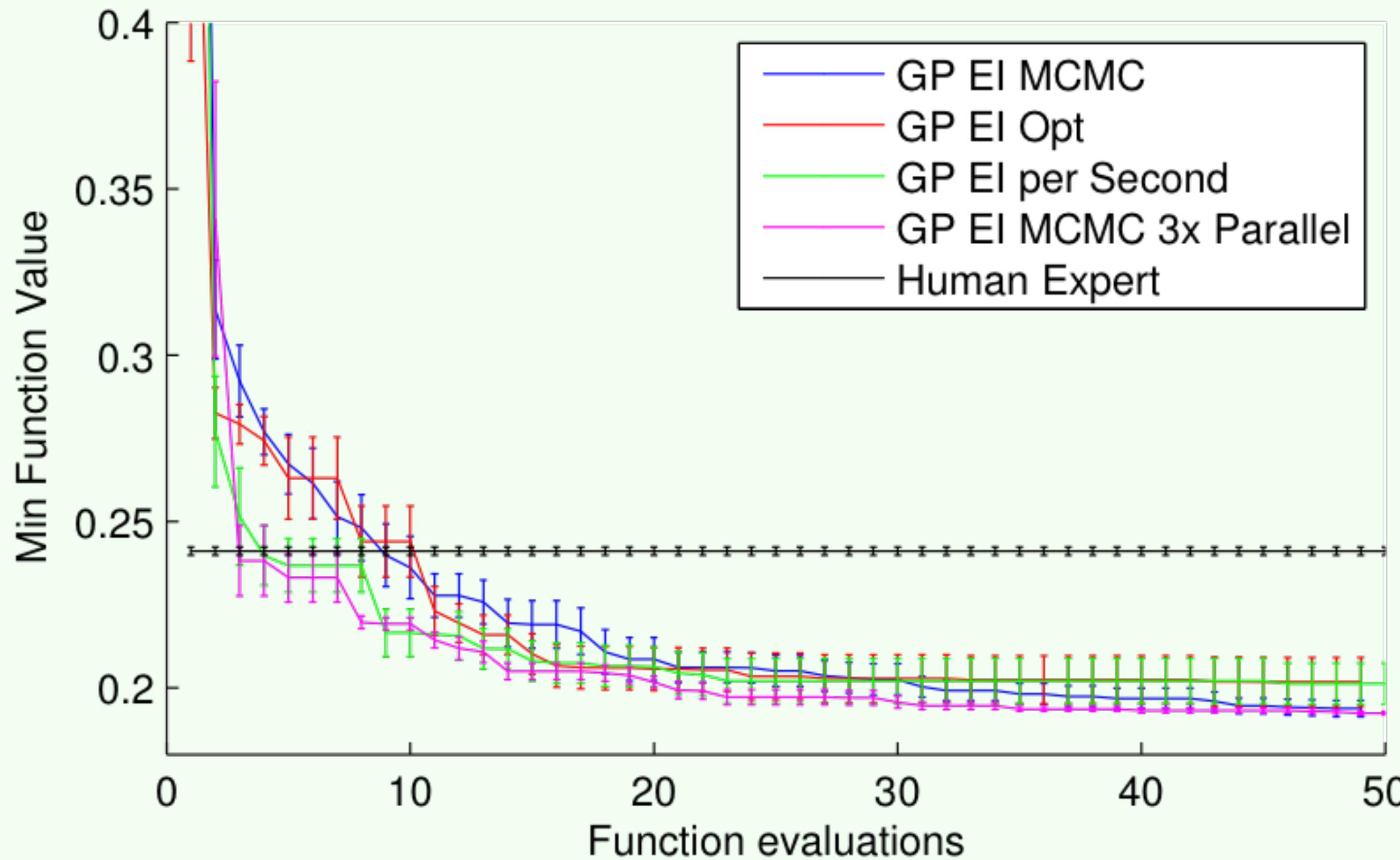
We make machine
learning better and
faster, automatically.



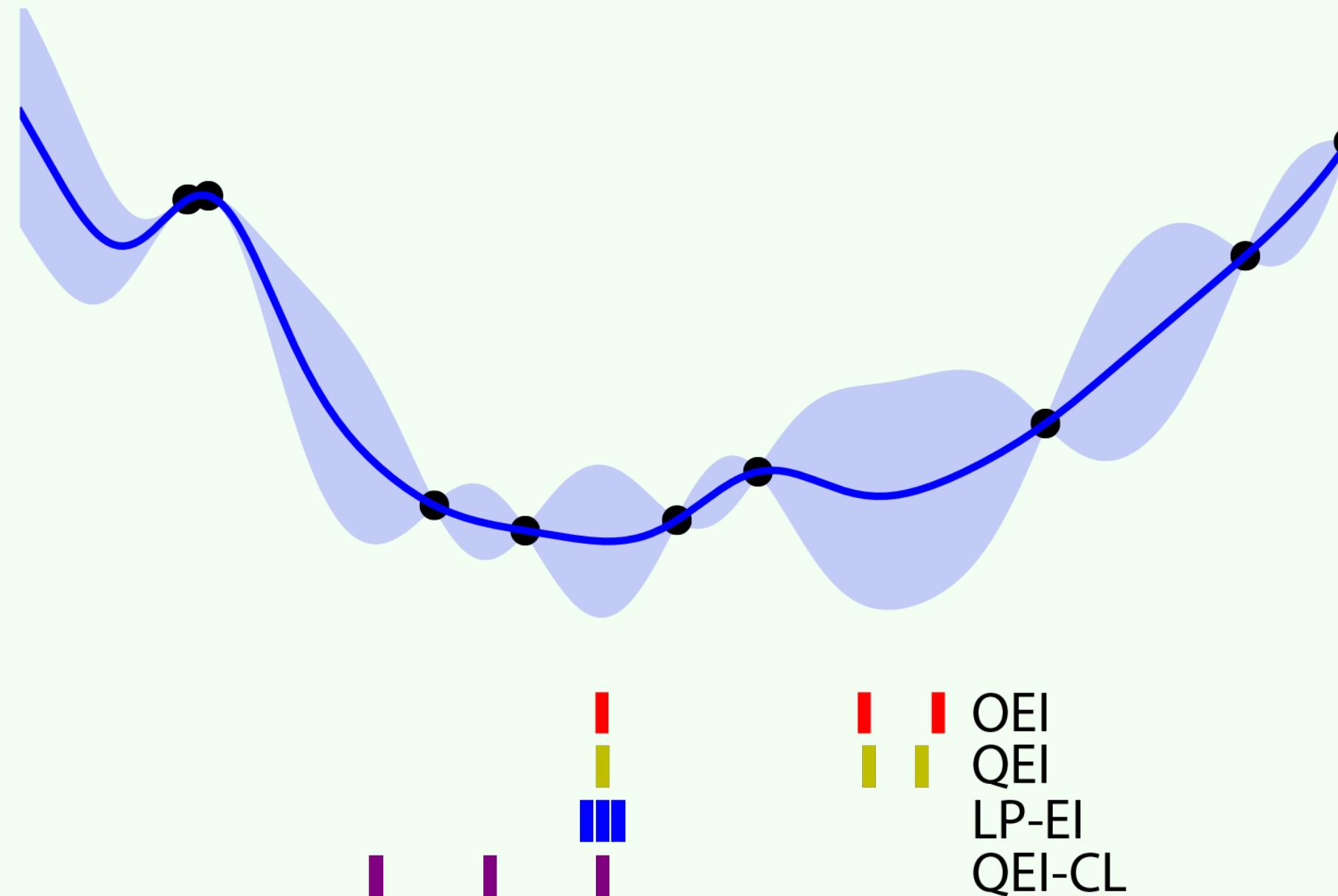
DataRobot



**MIND
FOUNDARY**
Applied Machine Learning



Batch Bayesian optimisation is run in parallel.



```
ea = params[0]
wa = params[2:3]
secw, sesw = np.sqrt(ea)*np.cos(wa), np.sqrt(ea)*np.sin(wa)
```

Hyperparameter optimisation is often treated as a black-box optimisation problem.

```
bx, by = af.baseline_m(df.time, pv_base, df.px, df.py, nthr)
px, py = af.am_model_em(df.time, np.r_[pv_base, pv_1, pv_2], 2, 1)
mx, my = 1e6*(bx+px), 1e6*(by+py)
```

```
ea = params[0]
```

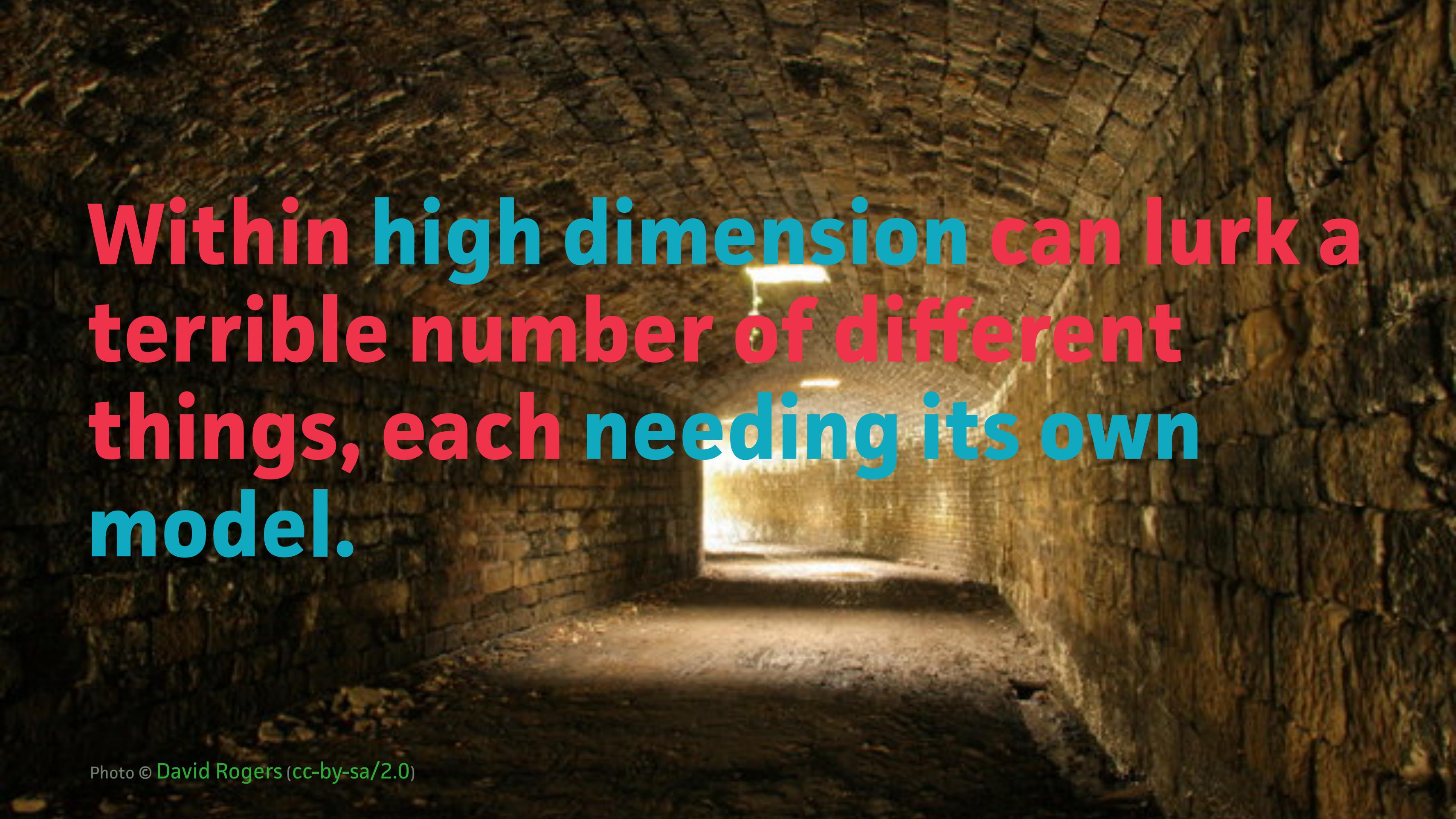
```
wa = params[2:3]
```

It is difficult to imagine a more white-box problem than one where you have full access to the problem's source code.

```
bx, by = af.baseline_m(df.time, pv_base, df.px, df.py, nthr)
```

```
px, py = af.am_model_em(df.time, np.r_[pv_base, pv_1, pv_2], 2, 1)
```

```
mx, my = 1e6*(bx+px), 1e6*(by+py)
```

A photograph of a dark, narrow stone tunnel. The walls are made of large, rough-hewn stones. Light from a single source at the far end creates a bright glow at the end of the tunnel, casting long shadows and illuminating the rough texture of the stone walls.

**Within high dimension can lurk a
terrible number of different
things, each needing its own
model.**

**Hyperparameters should usually
be marginalised, not optimised.**



Huge thanks to
Roman Garnett &
Philipp Hennig.