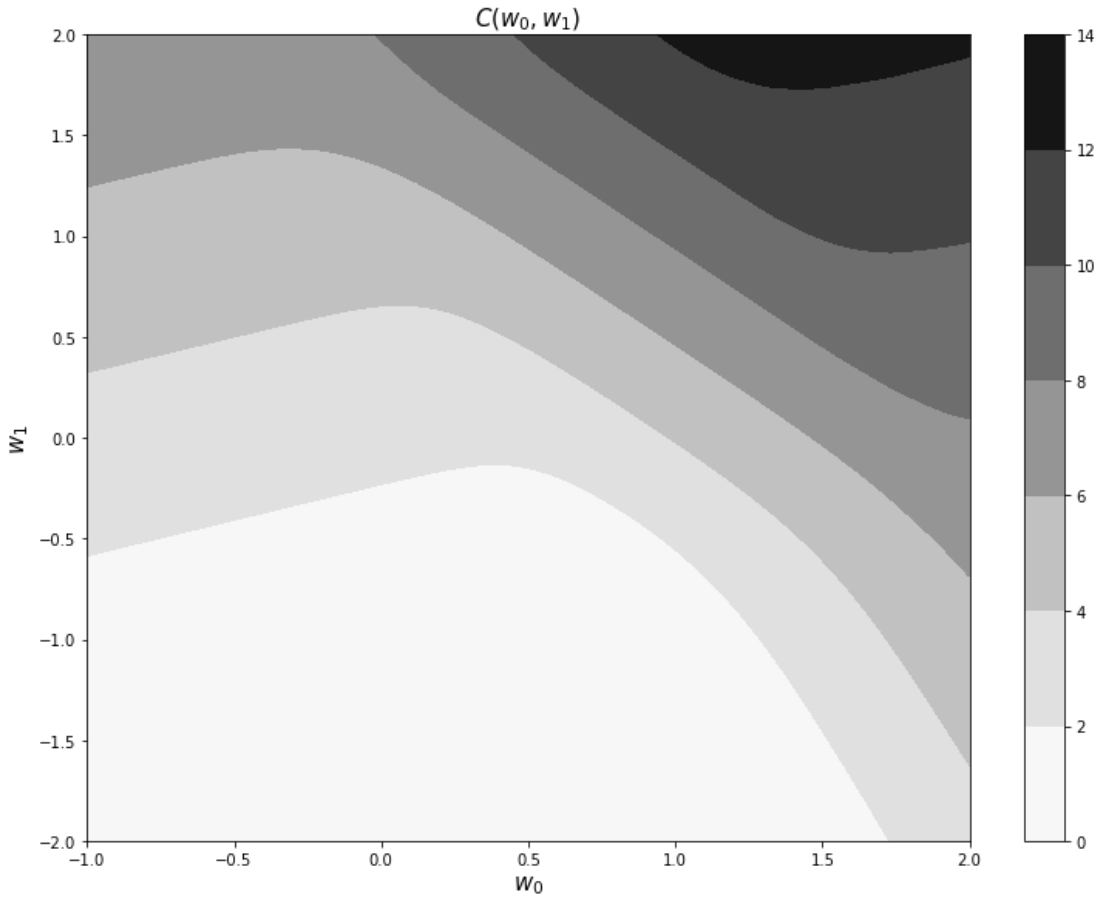
קצב הלמידה

בפרק זה נחזור למודל הנוירון היחיד אשר מסווג נקודות במרחב דו מימדי לשתי מחלקות – נקודות שחורות ונקודות לבנות. נזכור שהמודל תלוי בשלושה פרמטרים, w_0, w_1, b ומחזיר את הסתברות השייכות למחלקת הנקודות השחורות לפי הנוסחה

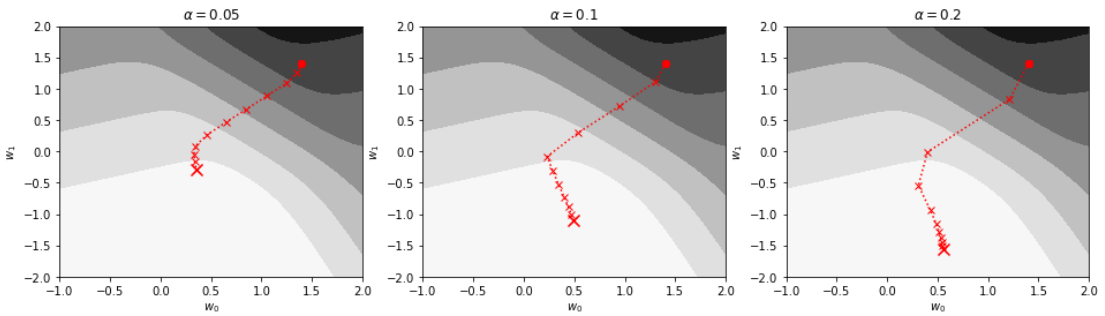
$$y = \frac{1}{1+e^{-(w_0x_0+w_1x_1+b)}}.$$

כעת נניח כי הפרמטר b קבוע מראש, ובעת אימון הנוירון אנו משנים את ערכי שני הפרמטרים האחרים. עבור בחירה מסויימת של פרמטרים אלו נקבל מודלים מוצלחים, ועבור בחירה אחרת – מודלים מוצלחים פחות, נצייר את מחיר האנטרופיה הצולבת המתקבל כתלות ב-w_0, w_1 , כאשר את ערכו של b קבענו על 5.

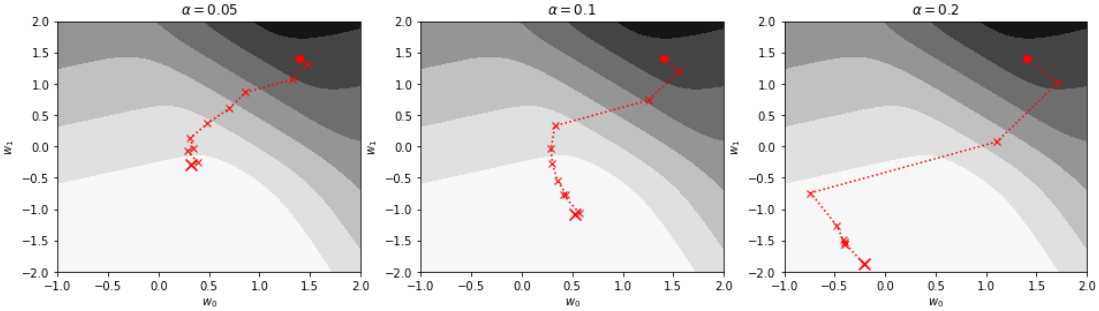


באיורים הבאים נדגים את תוצאות הפעלת אלגוריתם מורד הגרדיאנט (המלא), בו אנו מעדכנים את הפרמטרים לפי הנוסחה $(w_0, w_1) = (w_0, w_1) - \alpha \nabla C(w_0, w_1)$, עבור ערכי קצב למידה, α , שונים.

יש לשים לב שבכל אחת מהפעמים, איתחלנו את האלגוריתם עם הערכים $(w_0, w_1) = (1.4, 1.4)$, וכן שעשרת הצעדים הראשונים מסומנים ב-x.

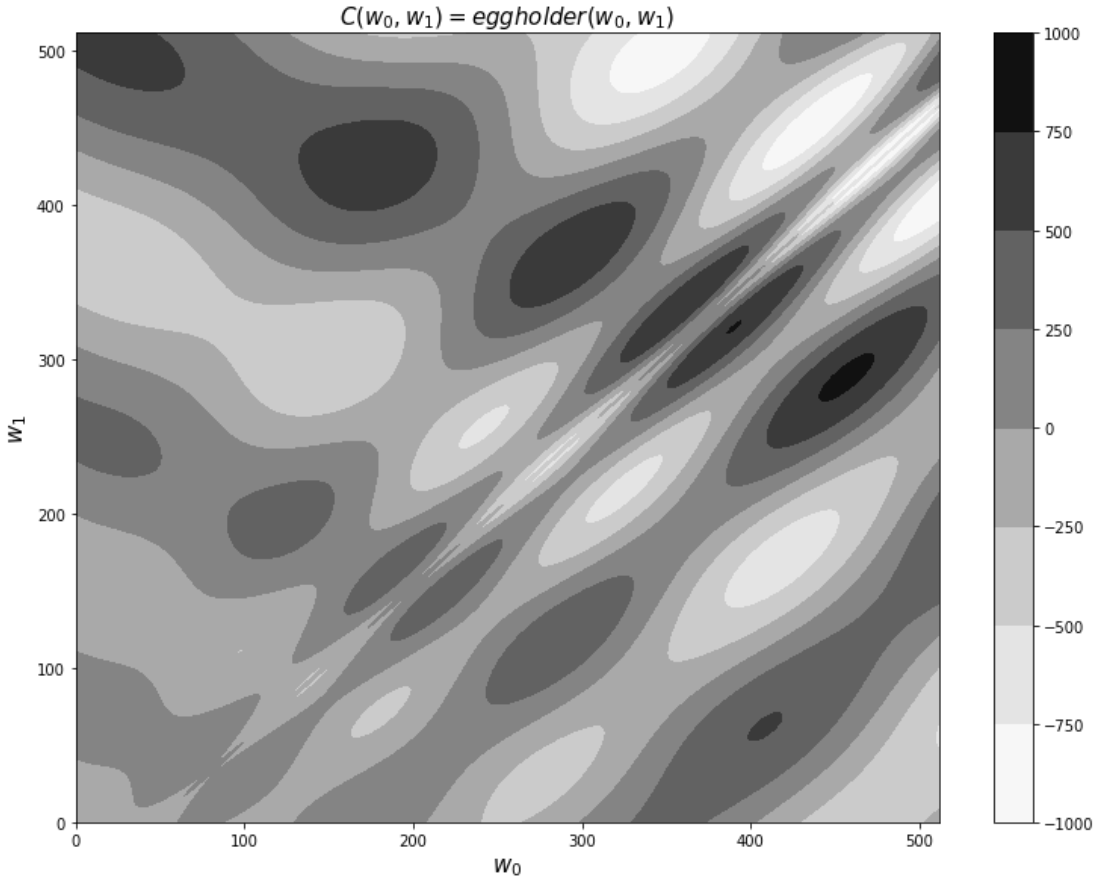


מאיורים אלו ניתן לראות את החשיבות של בחירת קצב הלמידה הנכון כבר בדוגמה פשוטה שכזו: קצב איטי מדי יוביל להתכנסות איטית. כאשר אנו משתמשים בגרדיאנט מקרי, חשיבות קצב הלמידה אף משמעותית יותר. נחליף את האלגוריתם בחישוב הנ"ל ל-SGD, כאשר כל איטרציה נחשב על פני minibatch של שתי דגימות ונקבל את התוצאות המופיעות באיור הבא.



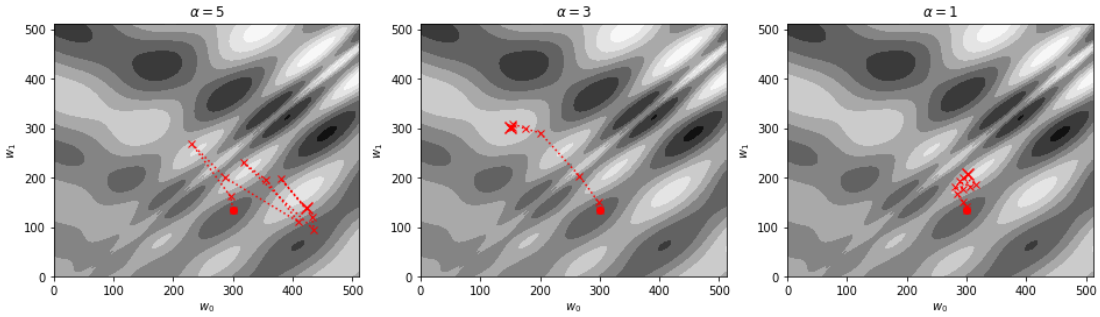
במעבר לגרדיאנט המקרי ניכר כי ההתכנסות איטית יותר, שכן לעתים הגרדיאנט המקרי רחוק מהגרדיאנט המלא, ובהתאם כיוון התנועה אינו אידיאלי, ומדוגמה זו ניכר כי פתרון אפשרי הוא להגדיל את קצב הלמידה. לרוב זהו אינו פתרון מתאים, שכן פונקציית המחיר של רשתות עמוקות יותר אינה נוחה לעבודה כמו הנ"ל, וקצב התכנסות גדול מדי עלול להוביל להתבדרות האלגוריתם. נניח לצורך הדיון כי אנו עוסקים במודל בעל שני פרמטרים, אך כעת פונקציית המחיר היא פונקציית "קרטון הביצים", אשר הגדרתה וציורה מופיעים להלן.

$$\text{eggholder}(w_0, w_1) = -(w_1 + 47)\sin\left(\sqrt{\left|w_1 + \frac{w_0}{2} + 47\right|}\right) - w_0 \sin\left(\sqrt{\left|w_0 - (w_1 + 47)\right|}\right)$$

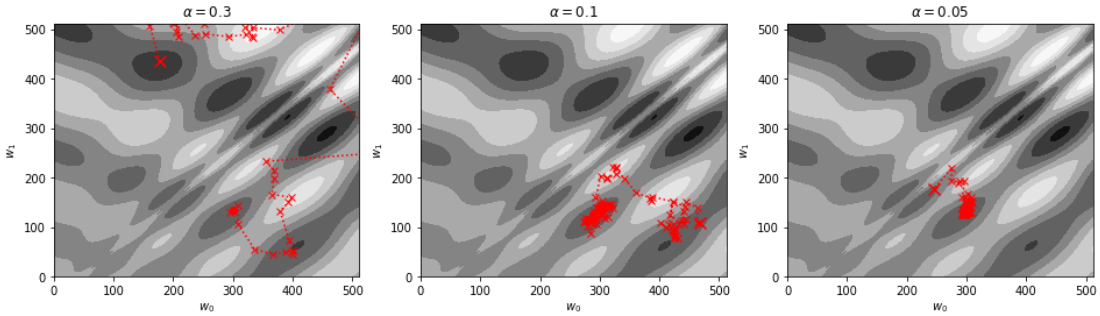


מציור הפונקציה ניתן להבין את משמעות שמה, והיא מהווה דוגמה מאתגרת לאלגוריתמי אופטימיזציה, שכן היא מכילה מספר רב של נקודות מינימום מקומיות.

תוצאות הרצת 100 איטרציות של אלגוריתם מורד הגרדיאנט (המלא) מאויירות להלן, ומהן ניכר שעבור קצב למידה גדול, האלגוריתם כבר אינו מתכנס.



המצב עבור מורד הגרדיאנט המקרי חמור אף יותר, כבר בקצבי למידה נמוכים כפי שמודגם באיור הבא.



שימו לב שאנו עוסקים כאן בסימולציה, שכן בדוגמה זו אין ברשותנו אוסף נתונים אשר ממנו אנו יכולים לדגום minibatch ועל בסיסה לחשב גרדיאנט מקרי, בהתאם אנו **מדמים** זאת על ידי חישוב הגרדיאנט האמיתי, והוספה של רעש רנדומלי, כפי שניתן לראות בשורת הקוד מלולאת האימון, בה אנו מעדכנים את ערכי הפרמטרים:

```
with torch.no_grad():
      w -= alpha*(w.grad*torch.normal(1,15,size=w.size()))
```

יחד עם זאת, דוגמאות אלו מייצגות נאמנה סיטואציות בהן ניתקל בעת אימון רשת עמוקה.
בהתבוננות באיורים הנ"ל ניתן לראות שעבור אף אחד מקצבי הלמידה הנבחרים, SGD לא התכנס כלל – ערכי הפרמטרים בהם עצר האלגוריתם רחוקים מלהיות מינימום מקומי, שכן הם אינם נמצאים במרכז "עמק" לבן. כלים לפתרון בעיה זו נלמד בהמשך היחידה הנוכחית.

שאלות לתרגול

1. לפונקציית קרטון הביצים נק' מינימום מקומי בקרבת הנק' $(w_0, w_1) = (400,130)$. נסו לאתר אותה בעזרת אלגוריתם מורד הגרדיאנט עם נקודות התחלה שונות וקצבי למידה שונים.
2. חזרו על פעולה זו עבור נק' המינימום המקומי הקרובה ל-$(w_0, w_1) = (450,420)$.
3. הסבירו למה קשה יותר למצוא את הנקודה השניה מאשר את הראשונה.