Mirror Descent Framework for Edge of Stability

Andrés Aradillas Fernández

August 2026

Abstract

We train a fully-connected multilayer perceptron with two hidden layers on the CIFAR-10 dataset. To train this model, we implement full-batch gradient descent (GD), stochastic gradient descent (SGD), Adam, and various other mirror descent optimizers, varying the learning rate to investigate the sensitivity of Edge of Stability (EoS) phenomena. We find limited evidence that mirror descent routines, when trained on the same dataset and optimizing the same neural architecture, trains at the EoS, while we are able to reproduce full-batch GD results from Cohen et al. (2021). Our results suggest non-GD mirror descent routines may not train at EoS as often as full-batch GD.

1 Introduction

Mirror descent is a more general framework that nests gradient descent. Previously, it has been shown that gradient descent methods, chief among them full-batch gradient descent, when applied to training neural networks, can produce a phenomenon known as "Edge of Stability" (henceforth EoS), first empirically described in detail in Cohen et al. (2021).

Sharpness is defined in the literature as the largest eigenvalue of the Hessian of the loss function evaluated at the current parameter iterate. Progressive sharpening, i.e., the continual increase early on in training of sharpness, is a well-documented phenomenon (see Jastrzebski et al. (2020)). More recently, EoS, the phenomenon where the sharpness hovers around a value of $2/\eta$, with η denoting the learning rate of a given optimization method, has been demonstrated to occur empirically, and there is a growing body of literature seeking to understand what drives this phenomenon, both from an empirical and theoretical framework.

2 Motivating Example

Since mirror descent nests gradient descent, and the intuition behind the latter's EoS phenomenon is cleaner, we can begin there. Gradient descent seeks to solve a convex loss function $L(\cdot)$ with solution $\theta \in \Theta$. If $\theta^\star$ represents the local minimum, then if $H_{\theta^\star} \equiv \nabla^2_\theta L(\theta^\star)$, then we can expand around $\theta^\star$:

$$L(\theta) \approx L(\theta^\star) + \frac{1}{2}(\theta - \theta^\star)^\top H_{\theta^\star}(\theta - \theta^\star).$$

Then, differentiating:

$$\nabla_\theta L(\theta) \approx H_{\theta^\star}(\theta - \theta^\star).$$

Therefore, as stated in Bubeck (2015), (full-batch) gradient descent takes the form

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t).$$

Substituting in our linear approximation of the subgradient, we get:

$$\theta_{t+1} = (I - \eta H_{\theta^\star})\theta_t$$

For stability, if $\rho(\cdot)$ denotes the spectral radius, i.e., $\rho(A) = \max_i |\lambda_i(A)|$, we will have stability if

$$\begin{aligned} &\rho(I - \eta H_{\theta^\star}) < 1 \\ \iff &|1 - \eta \lambda_{max}(H_{\theta^\star})| < 1 \\ \iff &0 < \eta \lambda_{max}(H_{\theta^\star}) < 2, \end{aligned}$$

and so we can see the edge of stability condition pops out at the boundary of the upper inequality:

$$\lambda_{max}(H_{\theta^\star}) < \frac{2}{\eta}.$$

Mirror descent operates in a more general way. It produces a sequence $\{\theta_t\}$ based on a mirror map $\Phi(\cdot)$. The intuition is that, by operating with this mirror map structure, mirror descent can generalize gradient descent methods for non-Euclidean geometries. In mirror descent, updates are produced according to

$$\nabla_\theta \Phi(\theta_{t+1}) = \nabla_\theta \Phi(\theta_t) - \eta \nabla L(\theta_t).$$

An analogous linearization around a local minimum $\theta^\star$ and defining $G(\theta) \equiv \nabla^2_\theta \Phi(\theta)$ then yields an approximation of the form

$$G(\theta^\star)\theta_{t+1} \approx G(\theta^\star)\theta_t - \eta H_{\theta^\star}\theta_t$$
$$\iff \theta_{t+1} \approx (I - \eta G(\theta^\star)^{-1} H_{\theta^\star})\theta_t.$$

Therefore, the more general edge of stability condition for mirror descent should resemble

$$\lambda_{max}(G^{-1}(\theta^\star)H_{\theta^\star}) < \frac{2}{\eta}.$$

A more computationally tractable quantity is

$$\lambda_{max}(G^{-1/2}(\theta^\star)H_{\theta^\star}G^{-1/2}(\theta^\star)).$$

We define $\text{EffectiveSharpness}_t$ as

$$\text{EffectiveSharpness}_t \equiv \lambda_{max}(G^{-1/2}(\theta_t)H_{\theta_t}G^{-1/2}(\theta_t)).$$

Computationally, since it is difficult to compute the exact set of eigenvalues, power iteration is used to compute the eigenvalues.

3 Experiments

3.1 Dataset

We trained a fully-connected ReLU network on CIFAR-10. The images in CIFAR-10 are flattened into 3072 input features. This setup is similar to the one considered in Cohen et al. (2021).

3.2 Network Architecture

The architecture of the network consists of two hidden layers with 512 and 256 as the corresponding widths. Each layer is initialized with Kaiming normal initialization and zero initial biases. The full CIFAR-10 dataset was used for training.

Batch size was 512. Sharpness was computed at every iteration. Power iteration was computed by iterating 10 times.

3.3 Optimizer Comparisons

3.3.1 Experiment 1

We first trained the multilayer perceptron (MLP) on full-batch gradient descent (GD), stochastic gradient descent (SGD) with a batch size of 512, signed entropy mirror descent, p-norm mirror descent, cosh mirror descent, Tsallis mirror descent, and AdaGrad mirror descent. Table 2 shows the learning rates corresponding to the optimizers.

Figure 1 shows EffectiveSharpness multiplied by the learning rate η across epochs. As can be seen, even with some mirror descent routines experiencing progressive sharpening, the only routine ultimately establishing itself into an EoS regime is full-batch gradient descent, despite all of these optimizers being used in the same setting on the same neural network architecture.

Table 1: Optimizers and Learning Rates for Experiment 1

Optimizer	**Learning Rate**
Gradient Descent (GD)	0.02
SGD	0.01
Entropy MD	0.1
p-norm MD ($p = 3$)	0.005
cosh MD	0.1
Tsallis MD	0.1
AdaGrad MD	0.3

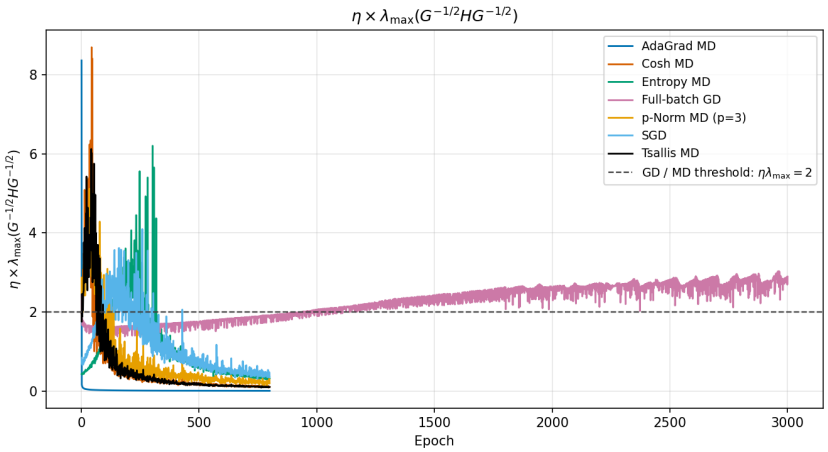


Figure 1: $\eta\times$EffectiveSharpness for Experiment 1

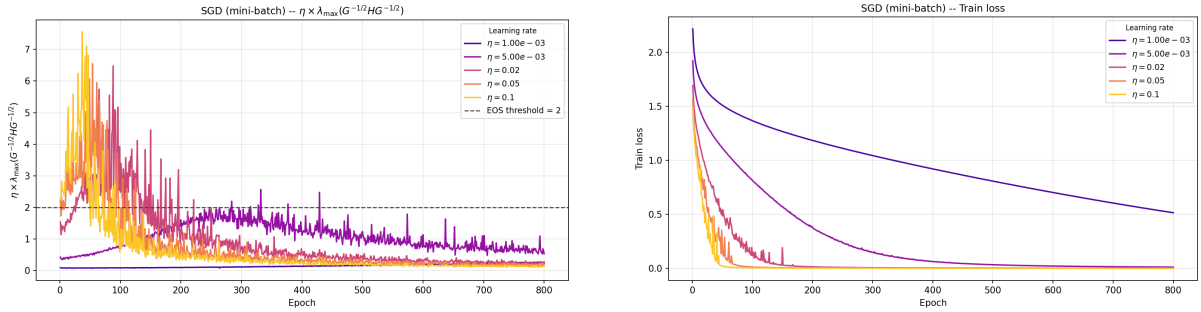
Table 2: Optimizers and Learning Rates for Experiment 2

Optimizer	**Learning Rate**
SGD	$\{0.001, 0.005, 0.02, 0.05, 0.1\}$
Entropy MD	$\{0.01, 0.03, 0.1, 0.3, 1.0\}$
p-norm MD ($p = 3$)	$\{0.001, 0.003, 0.01, 0.03, 0.1\}$

3.3.2 Experiment 2: Learning Rate Sensitivity

To further investigate whether the lack of EoS phenomenon is due to sensitivity of the learning rate parameter, we ran another experiment on a subset of the optimizers considered, running with various different learning rates. Table 2 shows the set of learning rates considered for each optimizer chosen.

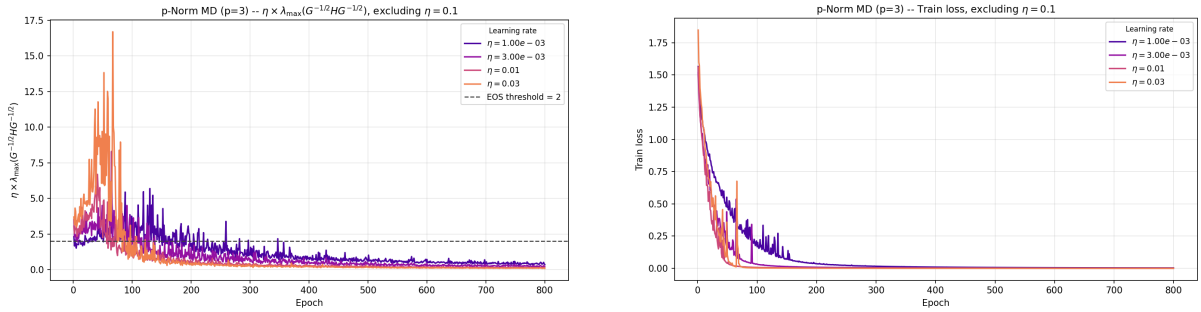
Figure 2a shows the evolution of effective sharpness across epochs for various different learning rates. While progressive sharpening is evident in the initial epochs for the larger learning rates, sharpness ultimately subsides, suggesting an approach to a (local) empirical risk minimizing (ERM) solution, irrespective of the learning rate. Loss also decreases over time nearly monotonically, as can be seen in Figure 2b.



(a) $\eta\times$ Effective Sharpness (b) Training Loss

Figure 2: Stochastic Gradient Descent in Experiment 2.

Figure 3a shows a similar figure but for p-norm mirror descent. The plot involving $\eta = 0.1$ was omitted, as it skewed the graph of the figure. The plot including $\eta = 0.1$ can be found in the Appendix in Figure 5. While the quantity of interest remains above the EoS threshold for $\eta = 0.1$, it does have volatile behavior around some threshold, suggesting EoS behavior may be occurring, but at a different threshold. However, this claim is speculative, and could also be a feature of the problem and associated loss space, rather than induced by some feature of p-norm mirror descent.



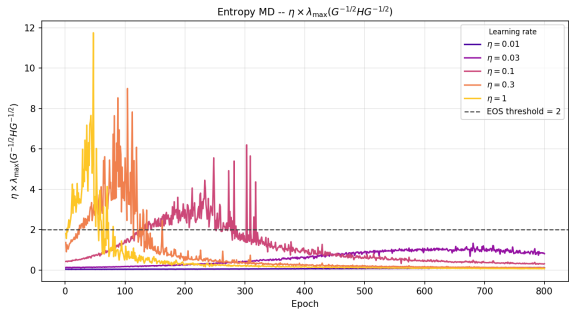
(a) $\eta\times$ Effective Sharpness (b) Training Loss

Figure 3: p-norm Mirror Descent in Experiment 2.

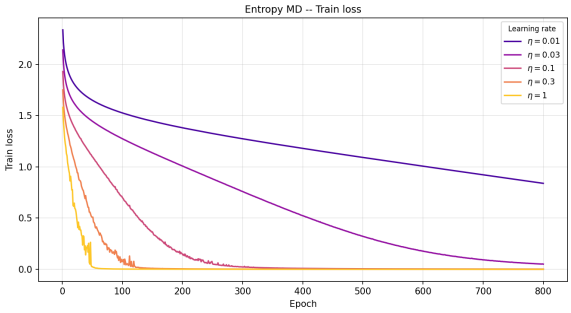
Lastly, Figure 4a shows the analogous figure for signed entropy mirror descent. Even as the learning rate is varied, while progressive sharpening evidently occurs for some learning rates, the effective sharpness eventually settles down, approaching zero. The training loss also decreases almost monotonically, as can be seen in Figure 4b.

4 Conclusion

While the architecture chosen was able to reproduce results from Cohen et al. (2021) that a fully-connected MLP trained on the CIFAR-10 dataset with full-batch gradient descent tends to train at the EoS, similar phenomena were not able to be induced for a variety of other optimizers: SGD, signed entropy MD, p-norm MD, cosh MD, Tsallis MD, and AdaGrad MD. Even when examining the sensitivity of the learning rates in training at EoS for SGD, signed entropy MD, and p-norm MD, progressive sharpening did generally



(a) $\eta\times$ Effective Sharpness



(b) Training Loss

Figure 4: Signed entropy mirror descent in Experiment 2.

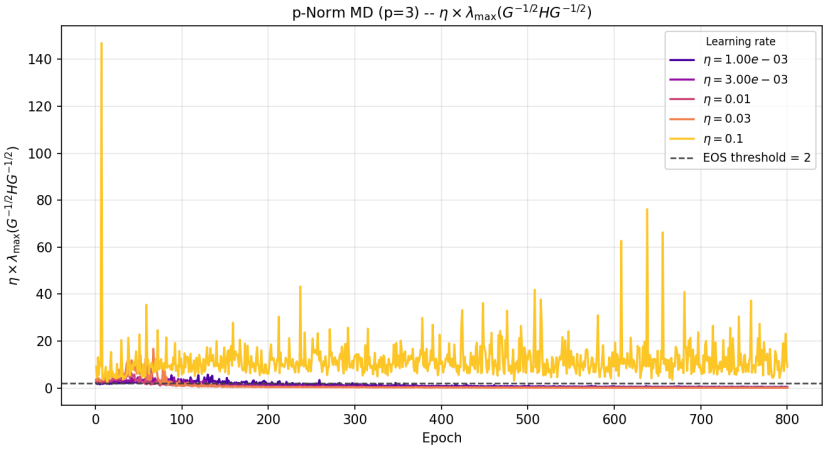


Figure 5: $\eta\times$EffectiveSharpness for p-norm MD in Experiment 2 with $\eta = 0.1$ included

occur, though very minimal training happened at the EoS. These results suggest that non-GD mirror descent routines may train less at EoS than full-batch GD, or at least such training does not necessarily happen when the MLP architecture is the same.

References

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and trends in Machine Learning*, 8(3-4):231–357.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. (2021). Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*.

Jastrzebski, S., Szymczak, M., Fort, S., Arpit, D., Tabor, J., Cho, K., and Geras, K. (2020). The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*.

Appendix

A Auxiliary Figures

A.1 Learning-Rate Sweeps

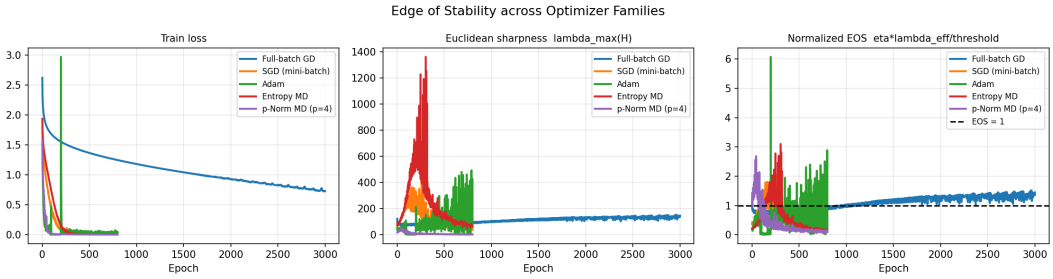
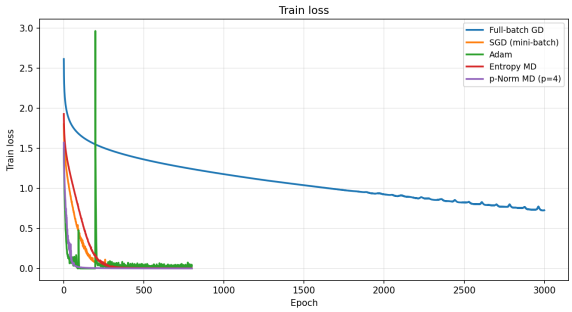
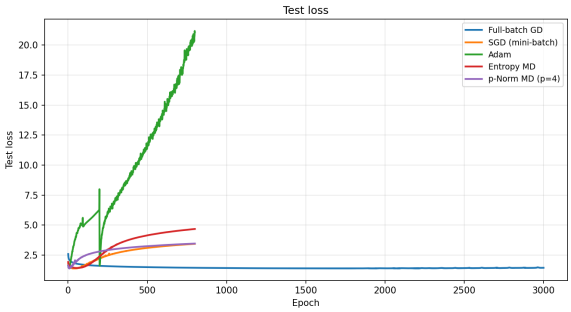


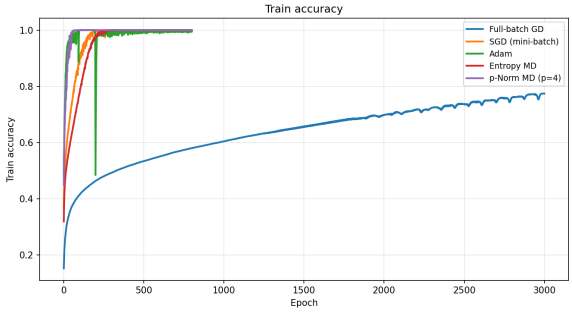
Figure 6: Main optimizer comparison, including full-batch GD, SGD, Adam, entropy mirror descent, and p-norm mirror descent: summary.



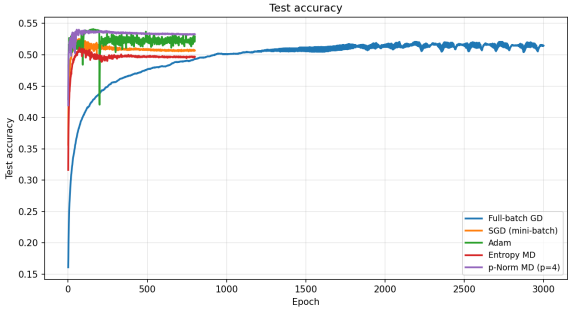
(a) Training loss



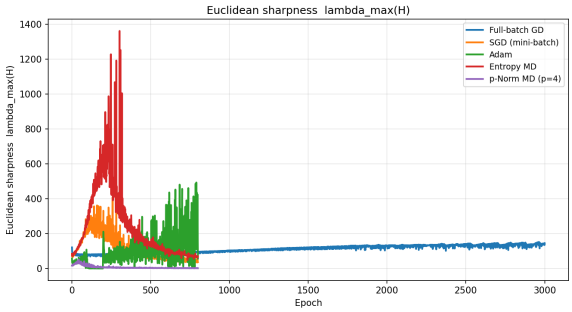
(b) Test loss



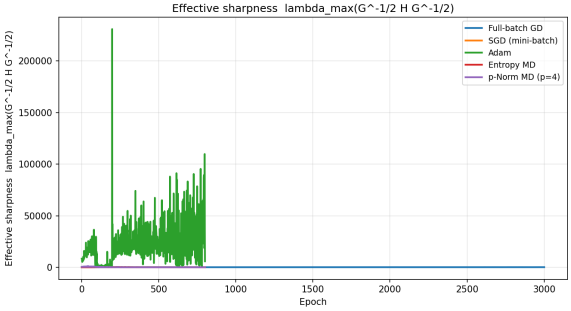
(c) Training accuracy



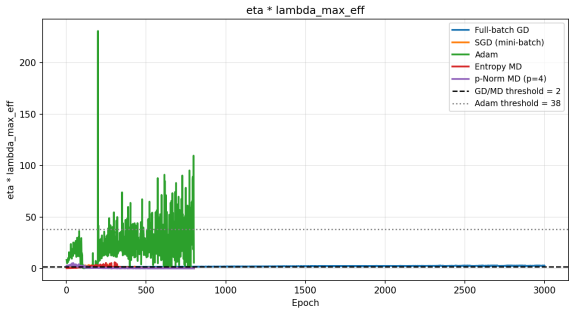
(d) Test accuracy



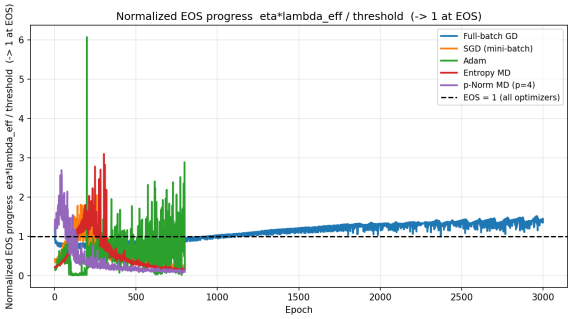
(e) $\lambda_{\max}(H)$



(f) $\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(g) $\eta\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(h) Normalized EOS

Figure 7: Main optimizer comparison, including full-batch GD, SGD, Adam, entropy mirror descent, and p-norm mirror descent: individual metric curves.

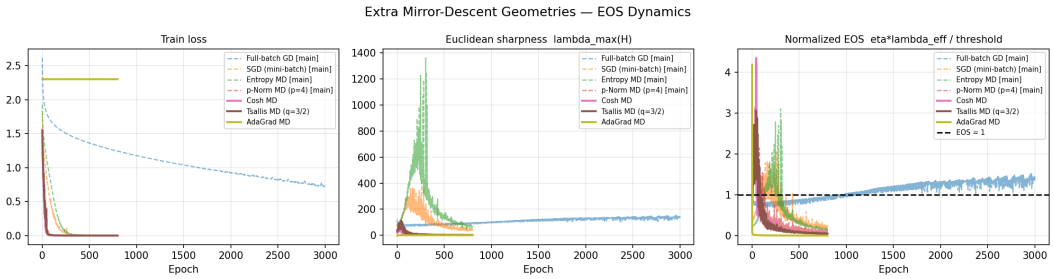
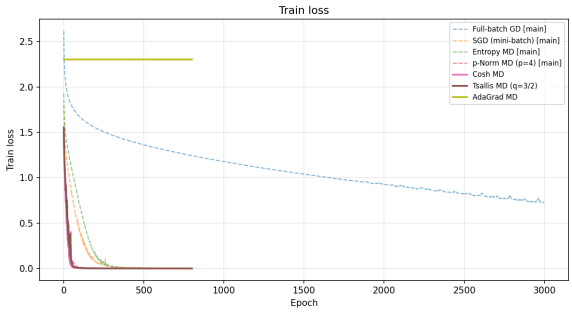
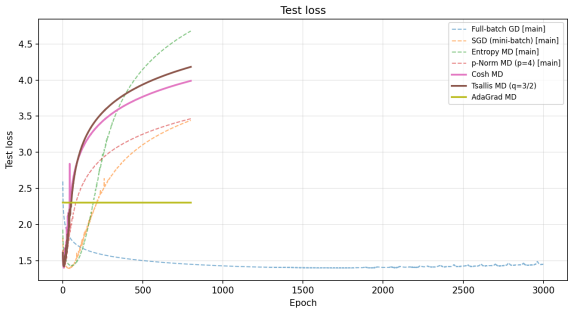


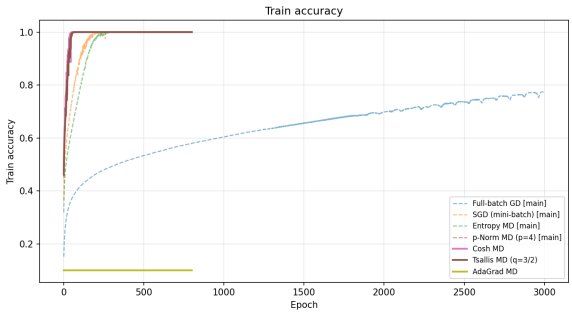
Figure 8: Additional mirror-descent geometries: summary.



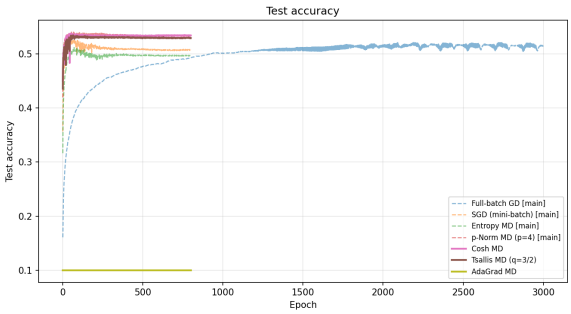
(a) Training loss



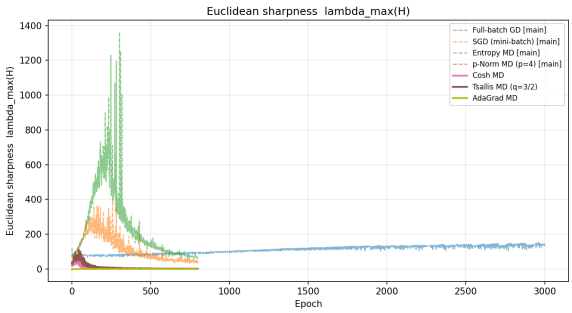
(b) Test loss



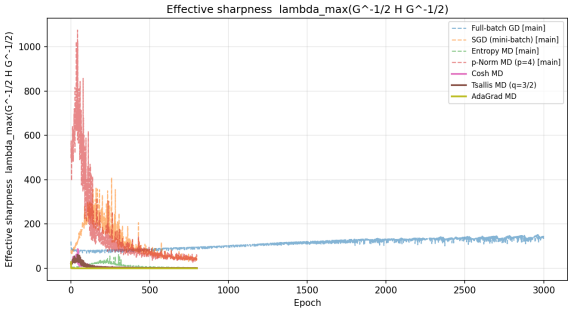
(c) Training accuracy



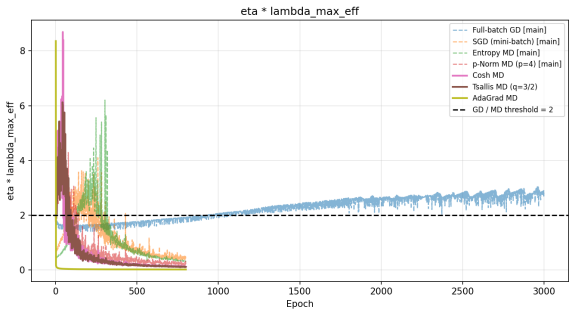
(d) Test accuracy



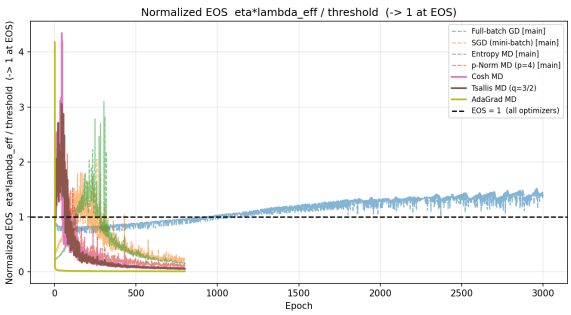
(e) $\lambda_{\max}(H)$



(f) $\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(g) $\eta\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(h) Normalized EOS

Figure 9: Additional mirror-descent geometries: individual metric curves.

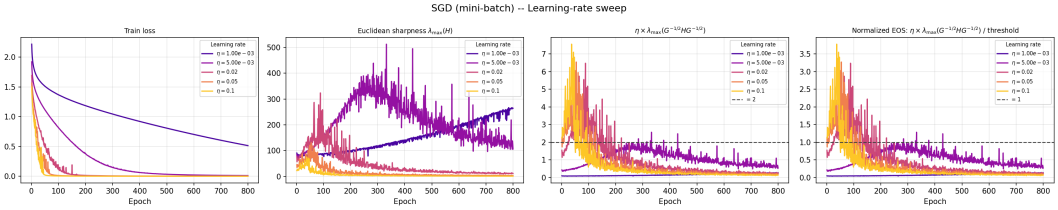
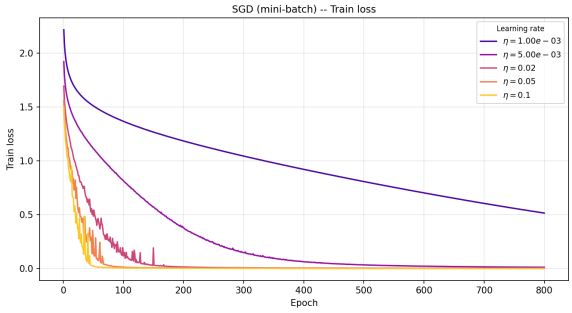
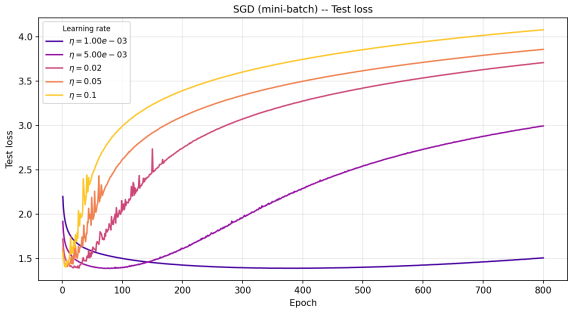


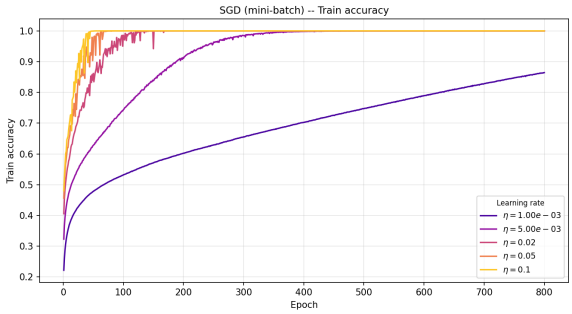
Figure 10: SGD learning-rate sweep: summary.



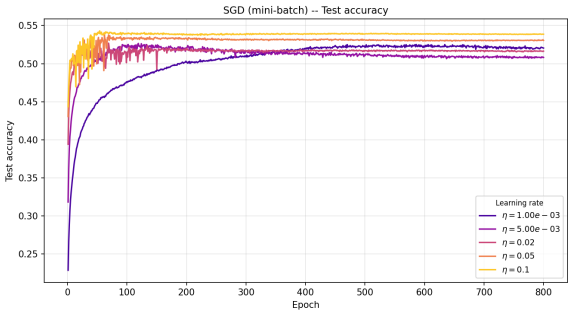
(a) Training loss



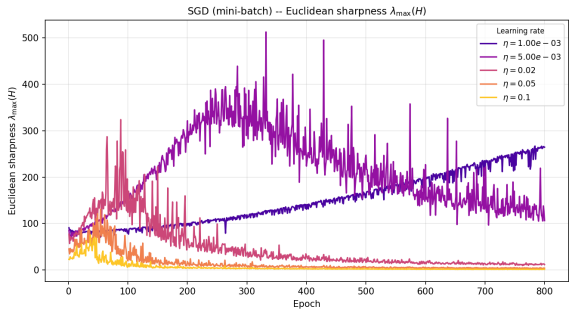
(b) Test loss



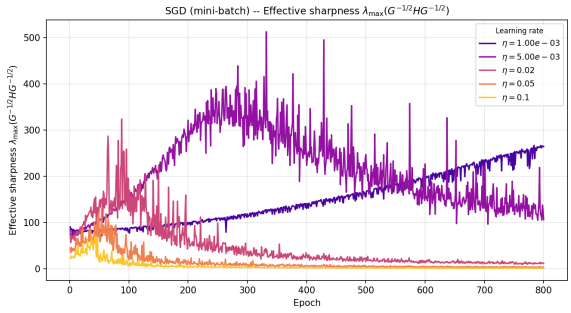
(c) Training accuracy



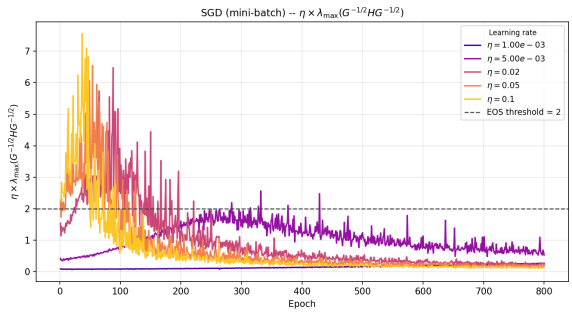
(d) Test accuracy



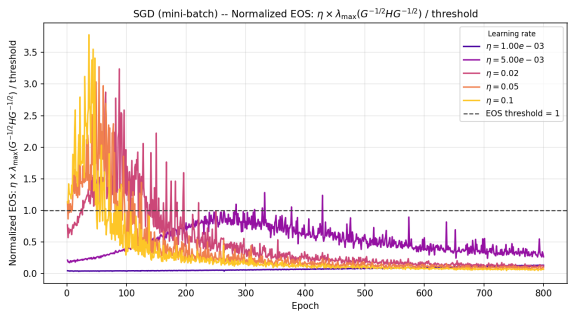
(e) $\lambda_{\max}(H)$



(f) $\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(g) $\eta\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(h) Normalized EOS

Figure 11: SGD learning-rate sweep: individual metric curves.

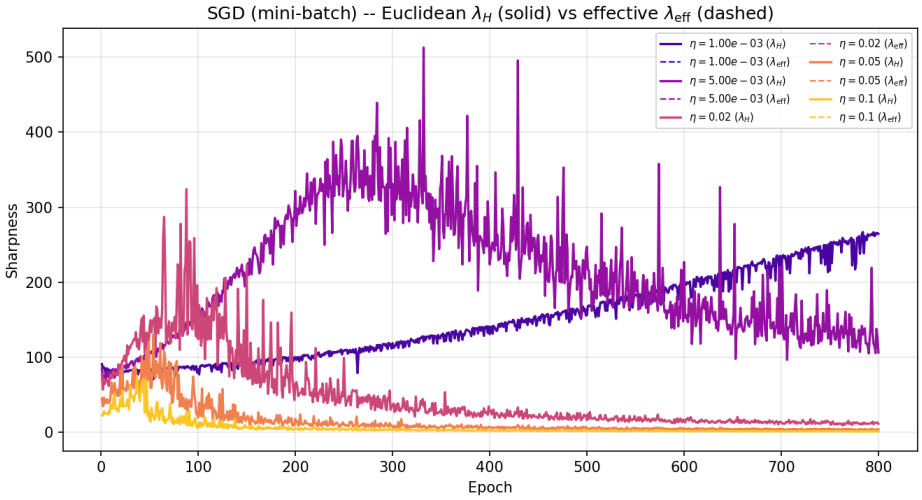


Figure 12: SGD learning-rate sweep: Euclidean sharpness λ_H and effective sharpness λ_{eff}.

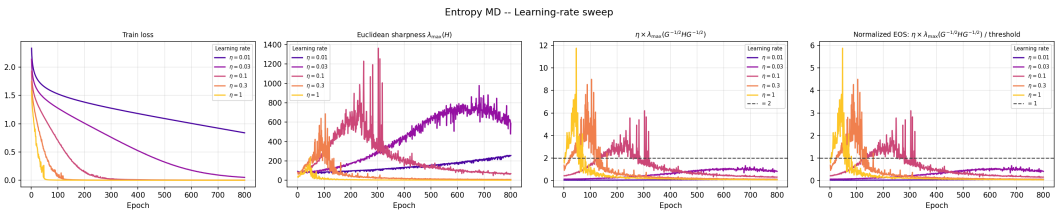
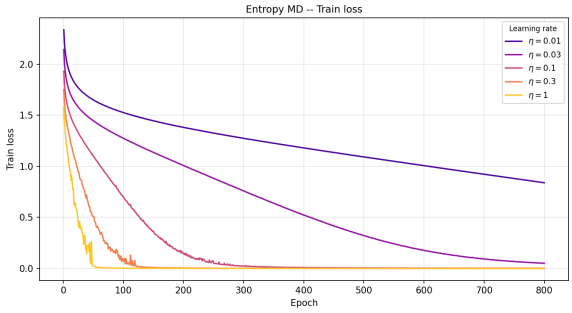
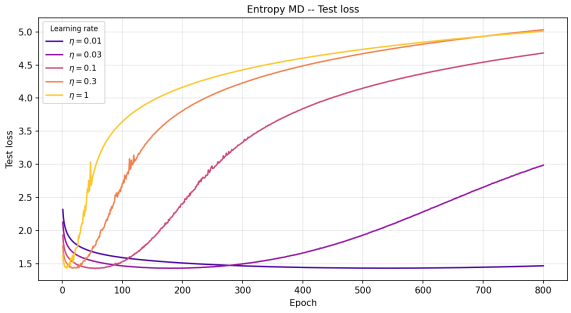


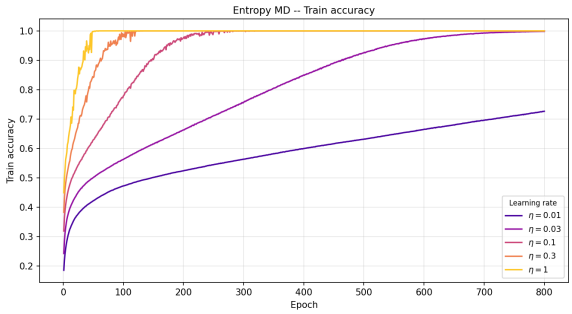
Figure 13: Entropy mirror descent learning-rate sweep: summary.



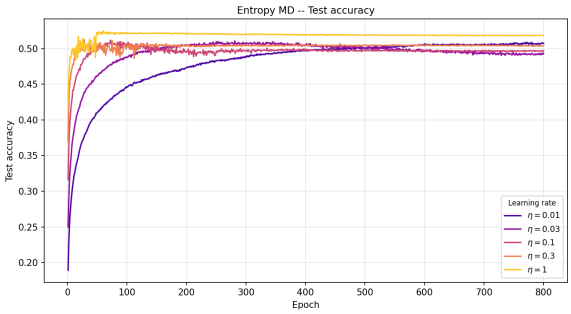
(a) Training loss



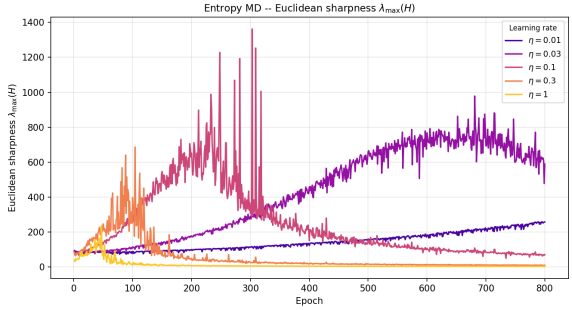
(b) Test loss



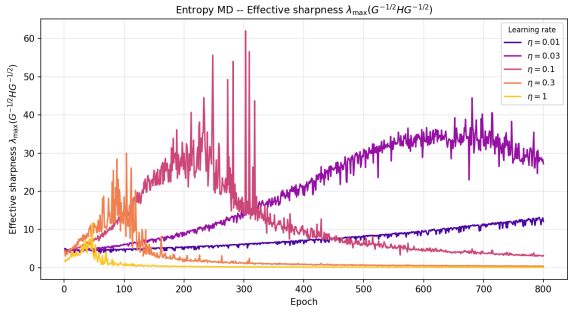
(c) Training accuracy



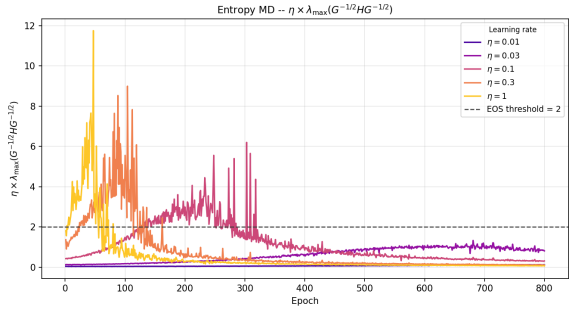
(d) Test accuracy



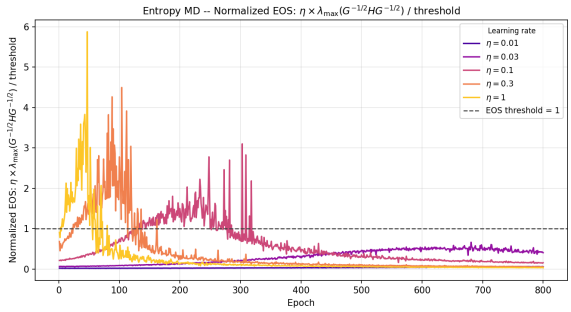
(e) $\lambda_{\max}(H)$



(f) $\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(g) $\eta\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(h) Normalized EOS

Figure 14: Entropy mirror descent learning-rate sweep: individual metric curves.

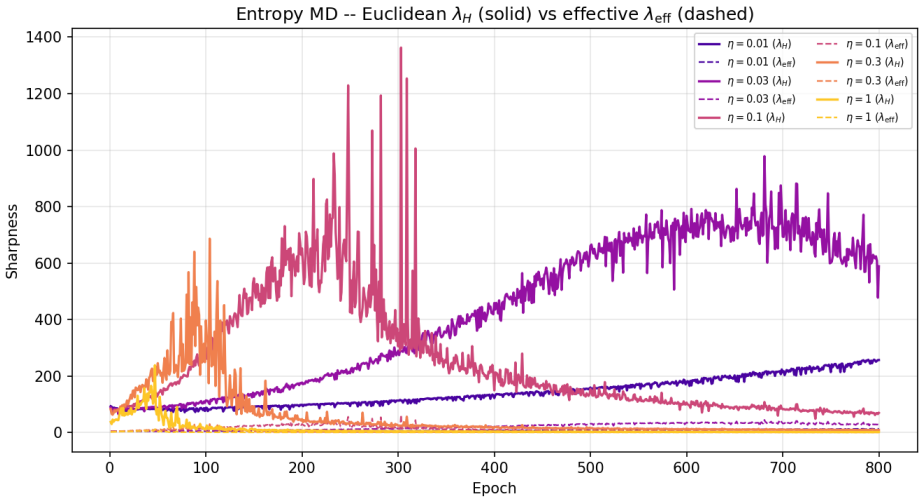


Figure 15: Entropy mirror descent learning-rate sweep: Euclidean sharpness λ_H and effective sharpness λ_{eff}.

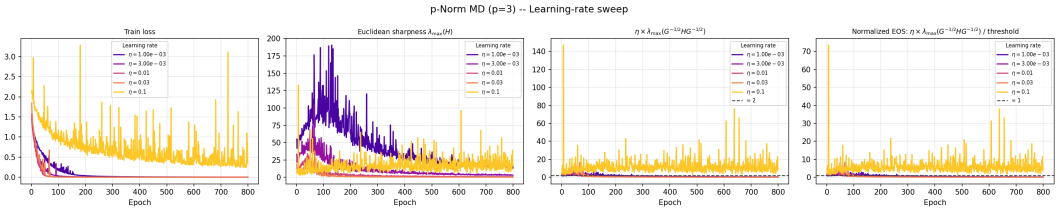
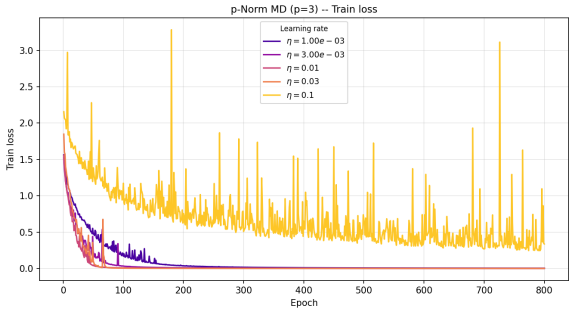
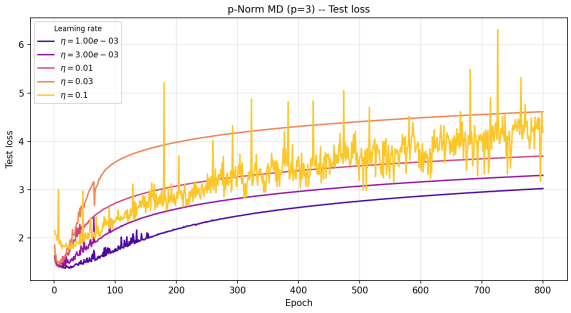


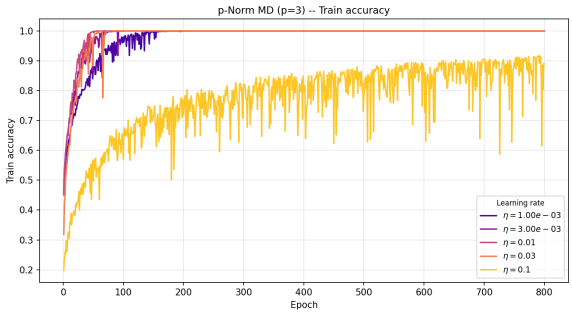
Figure 16: p-norm mirror descent with $p = 3$ learning-rate sweep: summary.



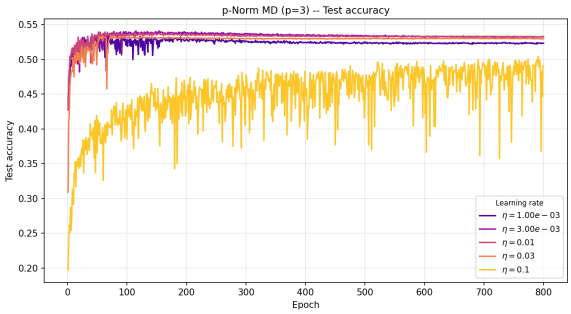
(a) Training loss



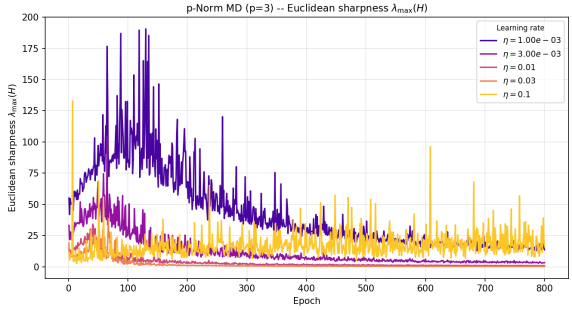
(b) Test loss



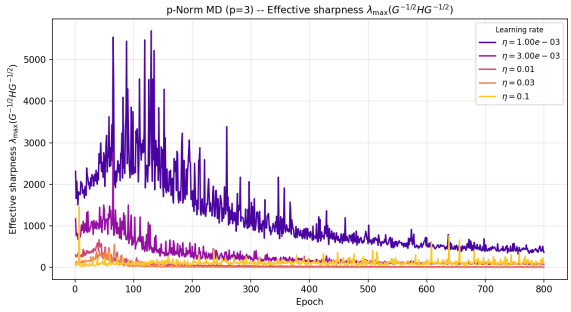
(c) Training accuracy



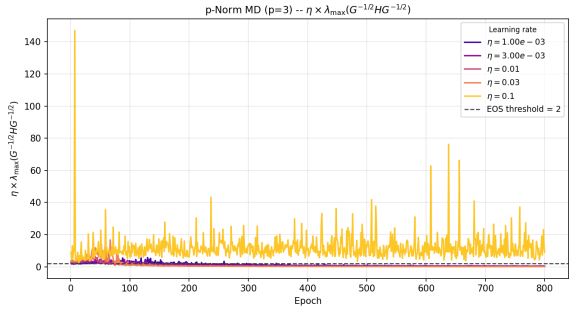
(d) Test accuracy



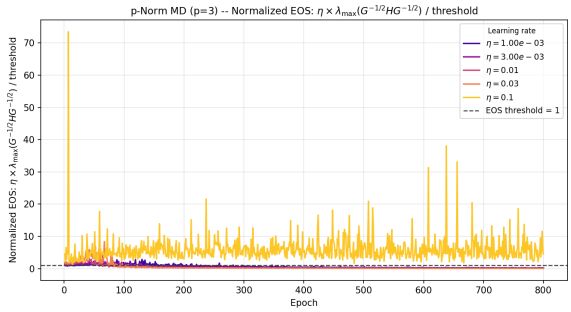
(e) $\lambda_{\max}(H)$



(f) $\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(g) $\eta\lambda_{\max}(G^{-1/2}HG^{-1/2})$



(h) Normalized EOS

Figure 17: p-norm mirror descent with $p = 3$ learning-rate sweep: individual metric curves.

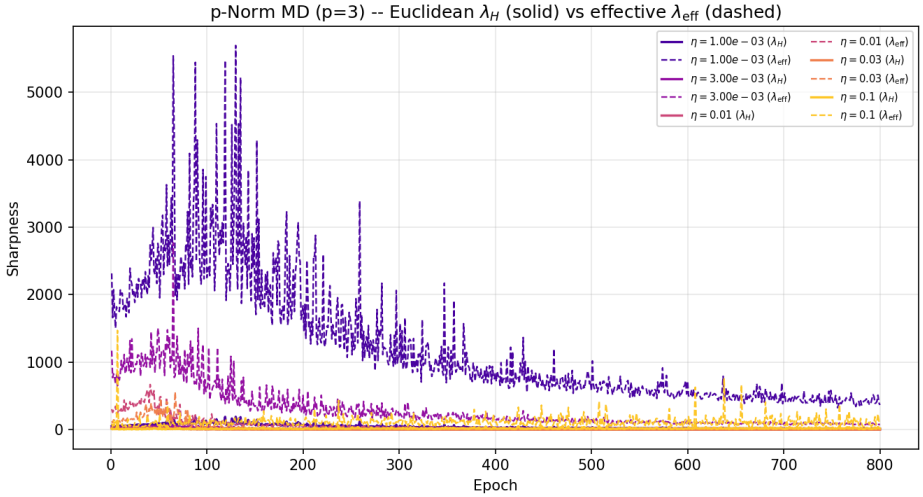
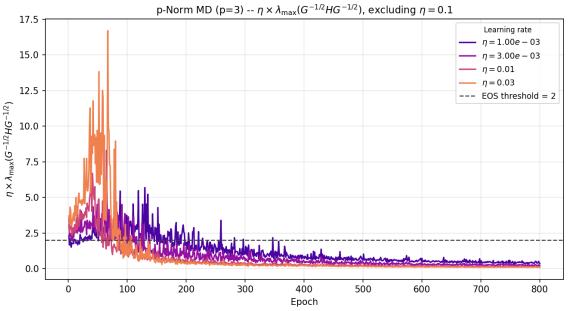
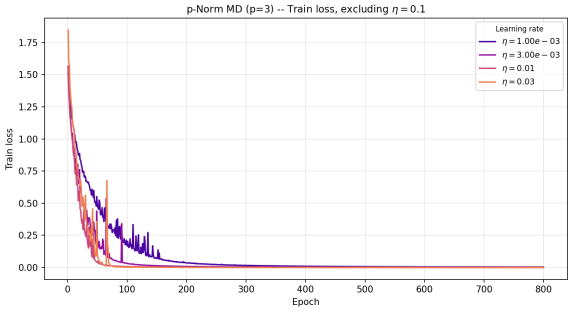


Figure 18: p-norm mirror descent with $p = 3$ learning-rate sweep: Euclidean sharpness λ_H and effective sharpness λ_{eff}.



(a) $\eta\lambda_{\max}(G^{-1/2}HG^{-1/2})$, excluding $\eta = 0.1$.



(b) Training loss, excluding $\eta = 0.1$.

Figure 19: Focused p-norm mirror descent sweep plots with the largest learning rate omitted. The p-norm runs use $p = 3$.