A Proofs

A.1 Bounds between H, F and C

A.1.1 Bounds with backward χ^2 divergence

$$\begin{aligned}
|\mathbf{F}_{ij} - \mathbf{H}_{ij}|^2 &= |\int q_\theta(x,y)\big(\nabla^2_\theta \ell(x,y)\big)_{ij} d(x,y) - \int p(x,y)\big(\nabla^2_\theta \ell(x,y)\big)_{ij} d(x,y)|^2 \\
&= |\int \big(q_\theta(x,y) - p(x,y)\big)\big(\nabla^2_\theta \ell(x,y)\big)_{ij} d(x,y)|^2 \\
&= |\int \frac{\big(q_\theta(x,y) - p(x,y)\big)}{\sqrt{p(x,y)}}\big(\sqrt{p(x,y)}\nabla^2_\theta \ell(x,y)\big)_{ij} d(x,y)|^2 \\
&\leq \int \frac{\big(q_\theta(x,y) - p(x,y)\big)^2}{p(x,y)} d(x,y) \int p(x,y)\big(\nabla^2_\theta \ell(x,y)\big)^2_{ij} d(x,y) \\
&= \mathcal{D}_{\chi^2}(q_\theta||p)\ \mathbb{E}_p[\big(\nabla^2_\theta \ell(x,y)\big)^2_{ij}]
\end{aligned}$$

Where we used Cauchy-Schwarz inequality and $\mathcal{D}_{\chi^2}$ denotes the χ^2 divergence.

$$||\mathbf{F} - \mathbf{H}||^2 \leq \mathcal{D}_{\chi^2}(q_\theta||p)\ \mathbb{E}_p[||\mathbf{H}(x,y)||_2^2]$$

Where $\mathbf{H}(x,y) \triangleq \nabla^2_\theta \ell(x,y)$ is the empirical hessian for one sample and the $||\cdot||_2$ is the Frobenius norm.

In the same way

$$\begin{aligned}
|\mathbf{F}_{ij} - \mathbf{C}_{ij}|^2 &= |\int q_\theta(x,y)\big(\nabla_\theta \ell(x,y) \nabla_\theta \ell(x,y)^\top\big)_{ij} d(x,y) - \int p(x,y)\big(\nabla_\theta \ell(x,y) \nabla_\theta \ell(x,y)^\top\big)_{ij} d(x,y)|^2 \\
&\leq \mathcal{D}_{\chi^2}(q_\theta||p)\ \mathbb{E}_p[\big(\nabla_\theta \ell(x,y) \nabla_\theta \ell(x,y)^\top\big)^2_{ij}]
\end{aligned}$$

For $\mathbf{C}(x,y) \triangleq \nabla_\theta \ell(x,y) \nabla_\theta \ell(x,y)^\top$ we have

$$||\mathbf{F} - \mathbf{C}||^2 \leq \mathcal{D}_{\chi^2}(q_\theta||p)\ \mathbb{E}_p[||\mathbf{C}(x,y)||^2]$$

Hence

$$||\mathbf{C} - \mathbf{H}||^2 \leq \mathcal{D}_{\chi^2}(q_\theta||p)\ \mathbb{E}_p[||\mathbf{C}(x,y)||^2 + ||\mathbf{H}(x,y)||^2]$$

A.1.2 Bounds with forward χ^2 divergence

Note that in the above proof, breaking the integral in two with Cauchy-Schwarz inequality could have been done using

$$\begin{aligned}
|\mathbf{F}_{ij} - \mathbf{H}_{ij}|^2 &= |\int \frac{\big(q_\theta(x,y) - p(x,y)\big)}{\sqrt{q_\theta(x,y)}}\big(\sqrt{q_\theta(x,y)}\nabla^2_\theta \ell(x,y)\big)_{ij} d(x,y)|^2 \\
&\leq \int \frac{\big(q_\theta(x,y) - p(x,y)\big)^2}{q_\theta(x,y)} d(x,y) \int q_\theta(x,y)\big(\nabla^2_\theta \ell(x,y)\big)^2_{ij} d(x,y) \\
&= \mathcal{D}_{\chi^2}(p||q_\theta)\ \mathbb{E}_{q_\theta}[\big(\nabla^2_\theta \ell(x,y)\big)^2_{ij}]
\end{aligned}$$

Similarly

$$|\mathbf{F}_{ij} - \mathbf{C}_{ij}|^2 \leq \mathcal{D}_{\chi^2}(p||q_\theta)\ \mathbb{E}_{q_\theta}[\big(\nabla_\theta \ell(x,y) \nabla_\theta \ell(x,y)^\top\big)^2_{ij}]$$

Thus

$$||\mathbf{C} - \mathbf{H}||^2 \leq \mathcal{D}_{\chi^2}(p||q_\theta)\ \mathbb{E}_{q_\theta}[||\mathbf{C}(x,y)||^2 + ||\mathbf{H}(x,y)||^2]$$

A.1.3 Proof of Proposition 2

From the upper bound assumption we have

$$\begin{aligned} f(\theta^{k+1}) &\leq f(\theta^k) + \nabla f(\theta^k)^\top(\theta^{k+1}-\theta^k) + \tfrac{1}{2}(\theta^{k+1}-\theta^k)^\top\mathbf{H}(\theta^{k+1}-\theta^k) \\ &= f(\theta^k) - \alpha\nabla f(\theta^k)^\top\mathbf{M}\nabla\ell(\theta^k,x) + \tfrac{\alpha^2}{2}\nabla\ell(\theta^k,x)^\top\mathbf{M}^\top\mathbf{H}\mathbf{M}\nabla\ell(\theta^k,x)\,. \end{aligned}$$

Subtracting $f(\theta^*)$ from both sides and taking conditional expectation we have

$$\begin{aligned} \mathbb{E}[f(\theta^{k+1}) - f(\hat{\theta}^*)] &\leq f(\theta^k) - f(\hat{\theta}^*) - \alpha\nabla f(\theta^k)^\top\mathbf{M}\mathbb{E}\big[\nabla\ell(\theta^k,x)\big] + \tfrac{\alpha^2}{2}\mathbb{E}\big[\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}\nabla\ell(\theta^k,x)\nabla\ell(\theta^k,x)^\top\right)\big] \\ &\leq f(\theta^k) - f(\hat{\theta}^*) - \alpha\nabla f(\theta^k)^\top\mathbf{M}\nabla f(\theta^k) + \tfrac{\alpha^2}{2}\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}(\mathbf{C}+\nabla f(\theta^k)\nabla f(\theta^k))\right) \\ &= f(\theta^k) - f(\hat{\theta}^*) - \alpha\nabla f(\theta^k)^\top(\mathbf{M} - \frac{\alpha}{2}\mathbf{M}^\top\mathbf{H}\mathbf{M})\nabla f(\theta^k) + \tfrac{\alpha^2}{2}\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}\mathbf{C}\right)\,, \end{aligned}$$

where in the second inequality we have used the covariance bound.

For $\mu_M\mathbf{I} \preccurlyeq \mathbf{M} - \frac{\alpha}{2}\mathbf{M}^\top\mathbf{H}\mathbf{M}$ and using the strong convexity bound $\frac{1}{2\mu}\|\nabla f(\theta)\|^2 \geq f(\theta) - f(\hat{\theta}^*)$, we can simplify to

$$\begin{aligned} \mathbb{E}[f(\theta^{k+1}) - f(\hat{\theta}^*)] &\leq f(\theta^k) - f(\hat{\theta}^*) - \alpha\mu_M\nabla f(\theta^k)^\top\nabla f(\theta^k) + \tfrac{\alpha^2}{2}\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}\mathbf{C}\right) \\ &\leq f(\theta^k) - f(\hat{\theta}^*) - 2\alpha\mu_M\mu\big(f(\theta^k) - f(\hat{\theta}^*)\big) + \tfrac{\alpha^2}{2}\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}\mathbf{C}\right) \\ &= \big(1-2\alpha\mu_M\mu\big)\big(f(\theta^k) - f(\hat{\theta}^*)\big) + \tfrac{\alpha^2}{2}\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}\mathbf{C}\right) \end{aligned}$$

Assuming $\alpha\mu_M\mu \leq \frac{1}{2}$, we have $\sum_{i=0}^{k}\left(1-2\alpha\mu_M\mu\right)^i \leq \sum_{i=0}^{\infty}\left(1-2\alpha\mu_M\mu\right)^i = \frac{1}{2\alpha\mu_M\mu}$. Therefore

$$\begin{aligned} \mathbb{E}[f(\theta^{k+1}) - f(\hat{\theta}^*)] &\leq f(\theta^k) - f(\hat{\theta}^*) - \alpha\mu_M\nabla f(\theta^k)^\top\nabla f(\theta^k) + \tfrac{\alpha^2}{2}\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}\mathbf{C}\right) \\ &\leq f(\theta^k) - f(\hat{\theta}^*) - 2\alpha\mu_M\mu\big(f(\theta^k) - f(\hat{\theta}^*)\big) + \tfrac{\alpha^2}{2}\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}\mathbf{C}\right) \\ &= \big(1-2\alpha\mu_M\mu\big)\big(f(\theta^k) - f(\hat{\theta}^*)\big) + \tfrac{\alpha^2}{2}\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}\mathbf{C}\right) \end{aligned}$$

Assuming $\alpha\mu_M\mu \leq \frac{1}{2}$, we have $\sum_{i=0}^{k}\left(1-2\alpha\mu_M\mu\right)^i \leq \sum_{i=0}^{\infty}\left(1-2\alpha\mu_M\mu\right)^i = \frac{1}{2\alpha\mu_M\mu}$. Taking full expectations and chaining inequalities we then have

$$\mathbb{E}[f(\theta^k) - f(\hat{\theta}^*)] \leq \big(1-2\alpha\mu_M\mu\big)^k\big(f(\theta^0) - f(\hat{\theta}^*)\big) + \tfrac{\alpha}{4\mu_M\mu}\operatorname{Tr}\left(\mathbf{M}^\top\mathbf{H}\mathbf{M}\mathbf{C}\right)\,.$$

This concludes the proof.

A.1.4 Convergence to limit cycles in the quadratic case

For SGD with constant stepsize α and preconditioner $\mathbf{M}$, the update equation on the parameters is

$$\theta_{t+1} = \theta_t - \alpha\mathbf{M}(\nabla f(\theta_t) + \epsilon_t)$$

In our quadratic case, $\nabla f(\theta_t) = \mathbf{H}(\theta_t - \theta^*)$ with $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\epsilon_t\epsilon_t^\top] = \mathbf{S}$. By defining $\delta_t = \mathbb{E}[\theta_t - \theta^*]$, we have

$$\begin{aligned} \delta_{t+1} &= (\mathbf{I} - \alpha\mathbf{M}\mathbf{H})\delta_t \\ &= (\mathbf{I} - \alpha\mathbf{M}\mathbf{H})^{t+1}\delta_0 \end{aligned}$$

This concludes the first result of proposition on the quadratic case.

By defining, $\mathbf{\Sigma}_t = \mathbb{E}[(\theta_t - \theta^*)(\theta_t - \theta^*)^\top]$, we get

$$\begin{aligned}
\mathbf{\Sigma}_{t+1} &= \mathbf{\Sigma}_t - \mathbb{E}\big[\alpha\mathbf{M}\big(\mathbf{H}(\theta_t - \theta^*) + \epsilon_t\big)(\theta_t - \theta^*)^\top\big] && (16)\\
&- \alpha\mathbb{E}\big[(\theta_t - \theta^*)\big(\theta_t - \theta^* + \epsilon_t\big)^\top\mathbf{H}\mathbf{M}^\top\big] && (17)\\
&+ \alpha^2\mathbb{E}\big[\mathbf{M}\mathbf{H}(\theta_t - \theta^*)(\theta_t - \theta^*)^\top\mathbf{H}\mathbf{M}^\top\big] && (18)\\
&+ \alpha^2\mathbb{E}\big[\mathbf{M}\epsilon_t\epsilon_t^\top\mathbf{M}^\top\big] && (19)\\
&= \mathbf{\Sigma}_t - \alpha\mathbf{M}\mathbf{H}\mathbf{\Sigma}_t - \alpha\mathbf{\Sigma}_t\mathbf{H}\mathbf{M}^\top + \alpha^2\mathbf{M}\mathbf{H}\mathbf{\Sigma}_t\mathbf{H}\mathbf{M}^\top + \alpha^2\mathbf{M}\mathbf{S}\mathbf{M}^\top && (20)\\
&= (\mathbf{I} - \alpha\mathbf{M}\mathbf{H})\mathbf{\Sigma}_t(\mathbf{I} - \alpha\mathbf{M}\mathbf{H})^\top + \alpha^2\mathbf{M}\mathbf{S}\mathbf{M}^\top && (21)\\
& && (22)
\end{aligned}$$

A.2 Expected suboptimality for SG and Polyak momentum on quadratic functions

We detail here the computation of the expected suboptimality at each timestep when optimizing a quadratic function with a diagonal Hessian when the noise is also diagonal. Note that all these results apply if $\mathbf{H}$ and $\mathbf{S}$ are simultaneously diagonalizable by a change of basis.

We assume that f is a quadratic with Hessian $\mathbf{H}$ and that, at each time step, we receive a gradient perturbed by a random variable ϵ with $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon\epsilon^\top] = \mathbf{S}$. Further, we shall assume that $\mathbf{H}$ and $\mathbf{S}$ are both diagonal. With these assumptions, the optimization occurs in each dimension independently and we can thus focus on a single dimension. We will denote by h and c the hessian and noise variance along that direction.

A.2.1 Proof of proposition 4

We can compare this result to the same setting where we use stochastic gradient with a diagonal preconditioning matrix $\mathbf{M}$. Then we get

$$\begin{aligned}
s_i &= (1 - \alpha\mathbf{M}_{ii}\mathbf{H}_{ii})^2 s_i + \alpha^2\mathbf{M}_{ii}^2\mathbf{S}_{ii}\\
s_i &= \frac{\alpha\mathbf{M}_{ii}\mathbf{S}_{ii}}{2\mathbf{H}_{ii} - \alpha\mathbf{M}_{ii}\mathbf{H}_{ii}^2}\,,
\end{aligned}$$

and

$$\mathbb{E}[f(\theta_t) - f(\hat{\theta}^*)] = \frac{1}{2}\sum_i \frac{\alpha\mathbf{M}_{ii}\mathbf{S}_{ii}}{2 - \alpha\mathbf{M}_{ii}\mathbf{H}_{ii}} + \mathcal{O}(e^{-t})\,.$$

Generalizing to simultaneously diagonalizable matrices, we get

$$\mathbb{E}[f(\theta_t) - f(\hat{\theta}^*)] = \frac{\alpha}{2}\operatorname{Tr}((2\mathbf{I} - \alpha\mathbf{M}\mathbf{H})^{-1}\mathbf{M}\mathbf{S}) + \mathcal{O}(e^{-t})\,.$$

A.2.2 Proof of proposition 5

Polyak momentum update equations are:

$$v_t = \gamma v_{t-1} + \nabla f(\theta_t) + \epsilon \quad (23)$$
$$\theta_{t+1} = \theta_t - \alpha v_t\,. \quad (24)$$

Using the quadratic assumption, we can rewrite

$$\begin{aligned}
v_{t+1} &= \gamma v_t + \nabla f(\theta_{t+1}) + \epsilon\\
&= \gamma v_t + h\theta_{t+1} + \epsilon\\
&= \gamma v_t + h\theta_t - \alpha h v_t + \epsilon\,,
\end{aligned}$$

and the full update can be written in matrix form

$$\begin{bmatrix} \theta_t \\ v_t \end{bmatrix} = \begin{bmatrix} 1 & -\alpha \\ h & \gamma - \alpha h \end{bmatrix}\begin{bmatrix} \theta_{t-1} \\ v_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ \epsilon \end{bmatrix} \quad (25)$$

Denoting $P = \begin{bmatrix} 1 & -\alpha \\ h & \gamma - \alpha h \end{bmatrix}$ and $S_t = [\begin{bmatrix} \theta_t \\ v_t \end{bmatrix} \begin{bmatrix} \theta_t \\ v_t \end{bmatrix}^T$, we have

$$\mathbb{E}[S_t | S_{t-1}] = P S_{t-1} P^T + \begin{bmatrix} 0 & 0 \\ 0 & c \end{bmatrix} . \tag{26}$$

If there is a limit cycle for $\begin{bmatrix} \theta_t \\ v_t \end{bmatrix}$, it will satisfy

$$S = P S P^T + \begin{bmatrix} 0 & 0 \\ 0 & c \end{bmatrix} . \tag{27}$$

Writing $S = \begin{bmatrix} s_\theta & s_{v\theta} \\ s_{v\theta} & s_v \end{bmatrix}$, we have

$$\begin{aligned} s_\theta &= s_\theta - 2\alpha s_{v\theta} + \alpha^2 s_v \\ s_v &= h^2 s_\theta + 2h(\gamma - \alpha h) s_{v\theta} + (\gamma - \alpha h)^2 s_v + c \\ s_{v\theta} &= h s_\theta + (\gamma - 2\alpha h) s_{v\theta} - \alpha(\gamma - \alpha h) s_v . \end{aligned}$$

The first equation gives $s_{v\theta} = \frac{\alpha}{2} s_v$ and the last one becomes

$$\begin{aligned} \frac{\alpha}{2} s_v &= h s_\theta + (\gamma - 2\alpha h)\frac{\alpha}{2} s_v - \alpha(\gamma - \alpha h) s_v \\ s_\theta &= \frac{\alpha(1+\gamma)}{2h} s_v . \end{aligned}$$

Finally, the second equation gives

$$\begin{aligned} s_v &= \left(h^2 \frac{\alpha(1+\gamma)}{2h} + 2h(\gamma - \alpha h)\frac{\alpha}{2} + (\gamma - \alpha h)^2 \right) s_v + c \\ s_v &= \frac{c}{(1-\gamma)\left(1 + \gamma - \frac{\alpha h}{2}\right)} \end{aligned}$$

and

$$s_\theta = \frac{\alpha(1+\gamma)c}{h(1-\gamma)(2+2\gamma-\alpha h)} .$$

Adding all dimensions together and multiplying by the Hessian to get the value function, we get

$$\mathbb{E}[f(\theta_t) - f(\hat{\theta}^*)] = \frac{1}{2} \sum_i \frac{\alpha(1+\gamma)\mathbf{S}_{ii}}{(1-\gamma)(2+2\gamma-\alpha\mathbf{H}_{ii})} + \mathcal{O}(e^{-t}) .$$

Generalizing to simultaneously diagonalizable matrices, we get

$$\mathbb{E}[f(\theta_t) - f(\hat{\theta}^*)] = \frac{\alpha}{2}\frac{(1+\gamma)}{(1-\gamma)} \operatorname{Tr}\left((2(1+\gamma)\mathbf{I} - \alpha\mathbf{H})^{-1}\mathbf{S}\right) + \mathcal{O}(e^{-t}) . \tag{28}$$

A.2.3 Comparison between stochastic gradient and Polyak momentum in the large noise regime

When the desired suboptimality is small, it requires a small α and the two suboptimality can be approximated by

$$f(\theta_t) - f(\hat{\theta}^*) \approx \frac{1}{4} \sum_i \frac{\alpha \mathbf{S}_{ii}}{(1-\gamma)} + o(1) \qquad \text{(Momentum)}$$

$$f(\theta_t) - f(\hat{\theta}^*) \approx \frac{1}{4} \sum_i \alpha \mathbf{S}_{ii} + o(1) , \qquad \text{(Stochastic gradient)}$$

and we see that momentum needs a stepsize α that is $(1-\gamma)$ times that of stochastic gradient to achieve the same suboptimality, countering any gain. This is what we see in Table 2.

B Experimental details

B.1 Details on the Hessian inverse

As $\mathbf{H}$ is highly degenerate in neural networks, we compute an inverse of $\mathbf{H}$ by cutting all the eigenvalues smaller than $10^{-3} \times \lambda_{max}$ where λ_{max} is the biggest eigenvalue of $\mathbf{H}$. We observed that 10^{-3} and 10^{-3} were reasonable constants for selecting the eigenvalues of significant magnitude. Using smaller constant sometimes lead to very noisy estimates of the TIC while using a bigger constant would lead to severe underestimation of the criterion.

B.2 Details on the large scale experiments

These details apply for the experiments conducted in subsection 5.5, figure 4 and all figures in subsection 5.1.

We remind the reader the setup.

- 5 different architectures: logistic regression, a 1-hidden layer and 2-hidden layer fully connected network, and 2 small convolutional neural networks (CNNs, one with batch normalization (Ioffe & Szegedy, 2015) and one without);
- 3 datasets: MNIST, CIFAR-10, SVHN;
- 3 learning rates: 10^{-2}, $5 \cdot 10^{-3}$, 10^{-3} using vanilla SGD with momentum $\mu = 0.9$;
- 2 batch sizes: 64, 512;
- 5 dataset sizes: 5k, 10k, 20k, 25k, 50k.

We train for 750k steps and compute our metrics every 75k steps.

Data preprocessing: We choose to greyscale, resize to 7×7 pixels and normalize all the images in the 3 datasets used (CIFAR-10, MNIST and SVHN). This way, we can design architectures with a relatively low number of parameters.

Architectures:

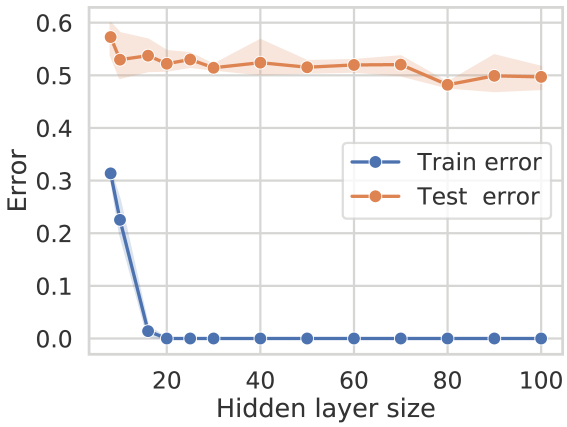
- `mlp`: This one is a one hidden layer MLP. Input size is $7 \times 7 = 49$ and output size is 10. The default number of hidden units is 70. We use ReLU activations.
- `big_mlp`: The architecture is the same as above but with one additional hidden layer.
- `logreg`: This is simple a 49×10 linear classifier.
- `cnn`: It is a small CNN with 3 layers. A first conv layer with kernel 3×3, 0 padding and 15 channels. The next layer has 20 channels and same parameters. The last layer has 10 channels and directly outputs the class scores.
- `cnn_bn`: Same architecture as above, except for a spatial batch-norm after the second layer.

B.3 Details on experiments of subsection 5.5

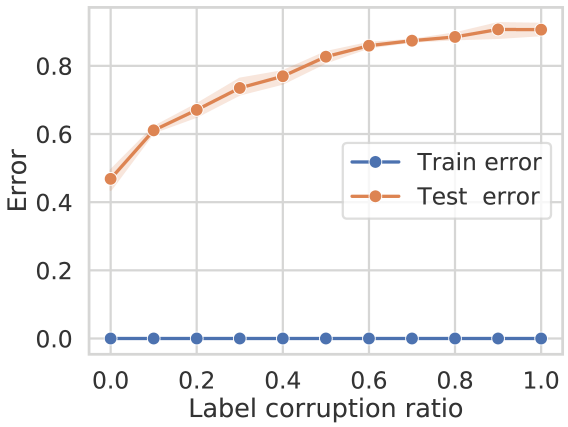
For these experiments we train one hidden layer MLPs on SVHN. Each points is computed by training three times with three different random seed until convergence. In figure 3a, the labels are kept without corruption and we vary the hidden size layer by using $\{8, 10, 16, 20, 25, 30, 40, 50, 60, 70, 80, 100\}$ hidden units in the hidden layer.

In figure 3b, we fix the number of hidden units to 70 but we vary the labels corruption percentage from 0% to 100% (included) by increments of 10%.

The networks are trained for $150k$ gradients steps with a learning rate of $5e{-}3$ and a batch size of 256. We used a subset of 2000 samples of SVHN to remain in the highly overparametrized regime, our networks were able to fit random data.



(a) Varying hidden layer size.



(b) Varying label randomization level.

Figure 6: The train and test errors associated with the experiments 3a and 3b. We see that while we use small networks, they are still able to fit the data completely provided we use more than 20 hidden units. This behavior mirrors the one of bigger networks.