# A SUPPLEMENTARY MATERIAL

## A.1 JD CONVERGENCE RATES

*Proof of Theorem 3.* Let $z = (x, y)$ be the location of our current iterate. Let $x^+ = x - (1/\beta)\nabla f(x)$ be our next iterate after a gradient step. By simply combining the $\beta$-smoothness with the definition of $x^+$, we have

$$\begin{aligned} f(x^+, y) &\leq f(x, y) - \frac{1}{\beta}\langle \nabla_x f(x, y), \nabla_x f(x, y)\rangle + \frac{\beta}{2}\|\nabla_x f(x, y)\|^2 \\ &\leq f(x, y) - \frac{1}{\beta}\|\nabla_x f(x, y)\|^2 \end{aligned}$$

For the descent guarantee of the gradientless step, we use a random direction and so let $u$ be a standard multivariate Gaussian. Then, $\langle \nabla_y f(x, y), u\rangle$ is a 1-D Gaussian with variance $\|\nabla_y f(x, y)\|^2$. Therefore, $\mathbb{E}\left[|\langle \nabla_y f(x, y), u\rangle|\right] = c\|\nabla_y f(x, y)\|$ for some dimension-independent constant $c$. Since $u$ is symmetric, with probability at least $0.5$, we have the following descent guarantee if we let $y^+ = y - hu$

$$\mathbb{E}[f(x, y^+)] \leq f(x, y) - h\mathbb{E}\left[|\langle \nabla_y f(x, y), u\rangle|\right] + \frac{\beta}{2}h^2\mathbb{E}\left[\|u\|^2\right] \leq f(x, y) - hc\|\nabla_y f(x, y)\| + \frac{\beta}{2}h^2 n_y$$

Therefore, if we choose $h = O(\frac{1}{n_y\beta}\|\nabla_y f(x, y)\|)$,

$$\mathbb{E}[f(x, y^+)] \leq f(x, y) - \frac{1}{\gamma}\|\nabla_y f(x, y)\|^2$$

where $\gamma = \Theta(n_y\beta)$. Note that since our gradientless step uses the binary radius search with minimum radius $r = \frac{\epsilon}{\sqrt{n_y}\beta}$ (see Theorem 13 of Golovin et al. (2019)), we can approximately find the optimal radius as long as $\|\nabla_y f(x, y)\| \geq \epsilon$, which allows us to deduce that our descent guarantee holds, up to constants, with only $O(\text{poly}\log(n/\epsilon))$ extra iterations.

By combining the two guarantees together, since $z^+ = (x^+, y)$ with probability $p$ and $z^+ = (x, y^+)$ with remaining probability,

$$\begin{aligned} \mathbb{E}[f(z^+)] &\leq f(x, y) - \frac{p}{\beta}\|\nabla_x f(x, y)\|^2 - \frac{1-p}{\gamma}\|\nabla_y f(x, y)\|^2 \\ &\leq f(x, y) - \frac{1}{2\gamma}(\|\nabla_x f(x, y)\|^2 + \|\nabla_y f(x, y)\|^2) \\ &\leq f(x, y) - \frac{1}{2\gamma}\|\nabla f(z)\|^2 \end{aligned}$$

Note that the second line follows since $0.5 \leq p \leq \frac{n_y}{n_y+1}$ and noting that $\frac{p}{\beta} \geq \frac{1}{2\gamma}$.

Finally we claim that by strong convexity $\|\nabla f(z)\|^2 \geq \alpha(f(z) - f(z^*))$. This holds since

$$f(z) - f(z^*) \leq \nabla f(z)^\top (z - z^*) - \frac{\alpha}{2}\|z - z^*\|^2 \leq \|\nabla f(z)\|\|z - z^*\| - \frac{\alpha}{2}\|z - z^*\|^2 \leq \frac{1}{2\alpha}\|\nabla f(z)\|^2$$

where the last line holds by AM-GM.

By combining the last two claims and applying standard calculations, we deduce our guarantee:

$$\begin{aligned} \mathbb{E}\left[f(z^+)\right] - f(z^*) &\leq f(z) - f(z^*) - \frac{\alpha}{2\gamma}\|\nabla f(z)\|^2 \\ &\leq \left(1 - \frac{1}{\Theta(\kappa n_y)}\right)(f(z) - f(z^*)) \end{aligned}$$

Note that if $\|\nabla_y f(x,y)\| \leq \epsilon$, then if $\|\nabla_x f(x,y)\| \geq \epsilon$, we can use the fact that $\|\nabla_x f(x,y)\| \geq \frac{1}{2}\|\nabla f(z)\|$ to get a similar descent guarantee. Otherwise, both gradients are small, then we see that $\|\nabla f(z)\| \leq 2\epsilon$ with strong convexity guarantees that $f(z) - f(z^*) = O(\epsilon^2)$. □

### A.2 AJD CONVERGENCE RATES

To achieve acceleration, we first show an useful lemma.

**Lemma 5.** *Let $z_k$ be the iterates of running Accelerated Joint Descent (Algorithm 2) and $z$ be any point. For all $k \geq 0$,*

$$\mathbb{E}\left[f_\eta(z_k)\right] - f_\eta(z) \leq \left(1 - \frac{1}{8\sqrt{\kappa}(n_y+4)}\right)^k \xi_0 + 8\eta^2\beta^2\sqrt{\kappa}(n_y+4)^2$$

*where $\xi_0 = \frac{\alpha}{2}\|z_0 - z\|^2 + f_\eta(z_0) - f_\eta(z)$.*

*Proof of Lemma 5.* Let $z_k = (x_k, y_k), v_k$ be generated after $k$ iterations. Then, we compute $w_k$ and generate the stochastic gradient $g_\eta(w_k) = 2g_\eta^{0.5}(w_k)$. First, from Lemma 5 of Nesterov & Spokoiny (2011), we can related the norm of $g_\mu(z)$ to its expectation. Specifically,

$$\begin{aligned}
\mathbb{E}\left[\|g_\mu(x)\|^2\right] &= \frac{4}{2}\|\nabla_x f(x,y)\|^2 + \frac{4}{2}\mathbb{E}\left[\left\|\frac{f(x,y+\eta u) - f(x,y)}{\eta}u\right\|^2\right] \\
&\leq 2\|\nabla_x f_\eta(x,y)\|^2 + 2\left[4(n_y+4)\|\nabla_y f_\eta(x,y)\|^2 + 3\eta^2\beta^2(n_y+4)^3\right] \\
&\leq 8(n_y+4)\|\nabla f_\eta(x,y)\|^2 + 6\eta^2\beta^2(n_y+4)^3
\end{aligned}$$

By smoothness,

$$f_\eta(z_{k+1}) \leq f_\eta(w_k) - h\nabla f_\eta(w_k)^\top g_\mu(w_k) + \beta\frac{h^2}{2}\|g_\mu(w_k)\|^2$$

By taking expectations,

$$\begin{aligned}
\mathbb{E}\left[f_\eta(z_{k+1})\right] &\leq f_\eta(w_k) - h\|\nabla f_\eta(w_k)\|^2 + \beta\frac{h^2}{2}\mathbb{E}\left[\|g_\mu(w_k)\|^2\right] \\
&\leq f_\eta(w_k) - h\frac{1}{8(n_y+4)}\left(\mathbb{E}\left[\|g_\mu(w_k)\|^2\right] - 6\eta^2\beta^2(n_y+4)^3\right) + \beta\frac{h^2}{2}\mathbb{E}\left[\|g_\mu(w_k)\|^2\right] \\
&= f_\eta(w_k) - \frac{1}{2}\theta\mathbb{E}\left[\|g_\mu(w_k)\|^2\right] + \delta_\eta
\end{aligned}$$

where $\delta_\eta \leq \eta^2\beta^2(n_y+4)$. Note the first line follows since $\mathbb{E}\left[g_\eta\right] = \nabla f_\eta$ and the second line follows from our derivations above and the third line follows from grouping terms and using $-h^2\beta + \beta h^2/2 = -\theta/2$ by definition.

For some $z = (x, y)$, we define our potential function to be:

$$\xi_{k+1}(z) = \frac{\alpha}{2}\|v_{k+1} - z\|^2 + f_\eta(z_{k+1}) - f_\eta(z)$$

By using the definition of $v_{k+1}$ and expanding, we get

$$\xi_{k+1}(z) = \frac{\alpha}{2}\|(1-a)v_k + aw_k - z\|^2 - \frac{\theta\alpha}{a}g_\eta(w_k)^\top\left[(1-a)v_k + aw_k - z\right] + \frac{\theta^2\alpha}{2a^2}\|g_\eta(w_k)\|^2 + f_\eta(z_{k+1}) - f_\eta(z)$$

Now by taking expectations,

$$\begin{aligned}
\mathbb{E}\left[\xi_{k+1}(z)\right] &\leq \frac{\alpha}{2}\|(1-a)v_k + aw_k - z\|^2 - a\nabla f_\eta(w_k)^\top\left[(1-a)v_k + aw_k - z\right] \\
&\quad + \frac{\theta}{2}\mathbb{E}\left[\|g_\eta(w_k)\|^2\right] + \mathbb{E}\left[f_\eta(z_{k+1})\right] - f_\eta(z) \\
&\leq \frac{\alpha}{2}\|(1-a)v_k + aw_k - z\|^2 - a\nabla f_\eta(w_k)^\top\left[(1-a)v_k + aw_k - z\right] \\
&\quad + f_\eta(w_k) - f_\eta(z) + \delta_\eta \\
&\leq \frac{\alpha}{2}\|(1-a)v_k + aw_k - z\|^2 + \delta_\eta \\
&\quad + f_\eta(w_k) + \nabla f_\eta(w_k)^\top\left[az + (1-a)z_k - w_k\right] - f_\eta(z) \\
&\leq \frac{\alpha}{2}\|(1-a)v_k + aw_k - z\|^2 + \delta_\eta \\
&\quad + (1-a)(f_\eta(z_k) - f_\eta(z)) - \frac{a\alpha}{2}\|z - w_k\|^2 \\
&\leq \frac{\alpha}{2}(1-a)\|v_k - z\|^2 + \frac{\alpha}{2}a\|w_k - z\|^2 + \delta_\eta \\
&\quad + (1-a)(f_\eta(z_k) - f_\eta(z)) - \frac{a\alpha}{2}\|z - w_k\|^2 \\
&= (1-a)\xi_k(z) + \delta_\eta
\end{aligned}$$

The first line follows since $\theta = a^2/\alpha$, the second line follows from our previous bound on $\mathbb{E}\left[f_\eta(z_{k+1})\right]$, the third line follows since $av_k = z_k - (1+a)w_k$, the fourth line follows by first separating our expression into linear combinations and then applying strong convexity, and the fifth line follows by convexity on the distance function. Finally, we get our result by definition of $\xi_k(z)$

Therefore, by using the tower property of expectations,

$$\mathbb{E}\left[\xi_k(z)\right] \leq (1-a)^k\xi_0(z) + \sum_{i=1}^{k}(1-a)^k\delta_\eta$$

We conclude by noting $a = \sqrt{\alpha\theta} = (8\sqrt{\kappa}(n_y+4))^{-1}$ and $\sum_i(1-a)^k \leq a^{-1} = 8\sqrt{\kappa}(n_y+4)$.

□

Finally, we proceed with the proof of the accelerated convergence rate.

*Proof of Theorem 4.* First, we claim that $|f_\eta(z) - f(z)| \leq \frac{\eta^2}{2}\beta n_y$. Note that this follows from a straightforward calculation:

$$|f_\eta(z) - f(z)| \leq \left|\int \eta\nabla_y f^\top u + \frac{\eta^2}{2}\beta\|u\|^2\frac{1}{P}e^{-\|u\|^2/2}\,du\right| \leq \frac{\eta^2}{2}\beta\frac{1}{P}\int\|u\|^2e^{-\|u\|^2/2}\,du$$

The claim follows by evaluating the variance integral to be equal to $n_y$.

Then, we simply combine our claim with Lemma 5 to derive our accelerated theorem and noting that $\eta^2\beta n_y \leq \eta^2\beta^2\sqrt{\kappa}(n_y+4)^4$

□

### A.3 EXPERIMENT PLOTS

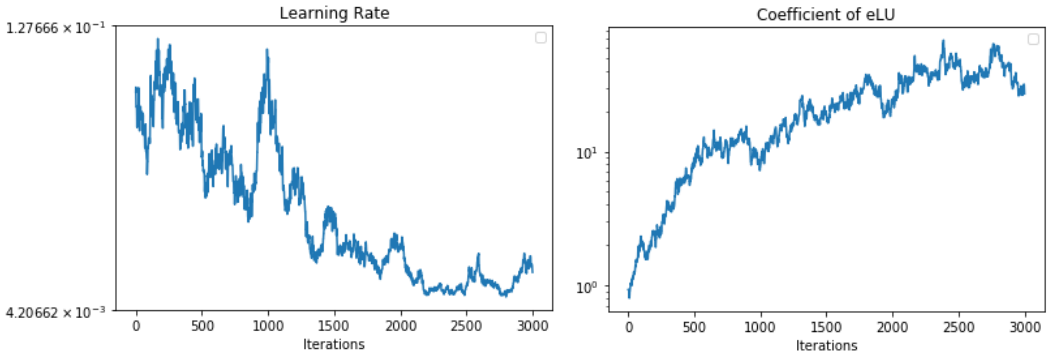


Figure 4: Convergence plot for Joint Descent for the MNIST dataset for the learning rate and the eLU coefficient. Note that the learning rate decreases as training converges.

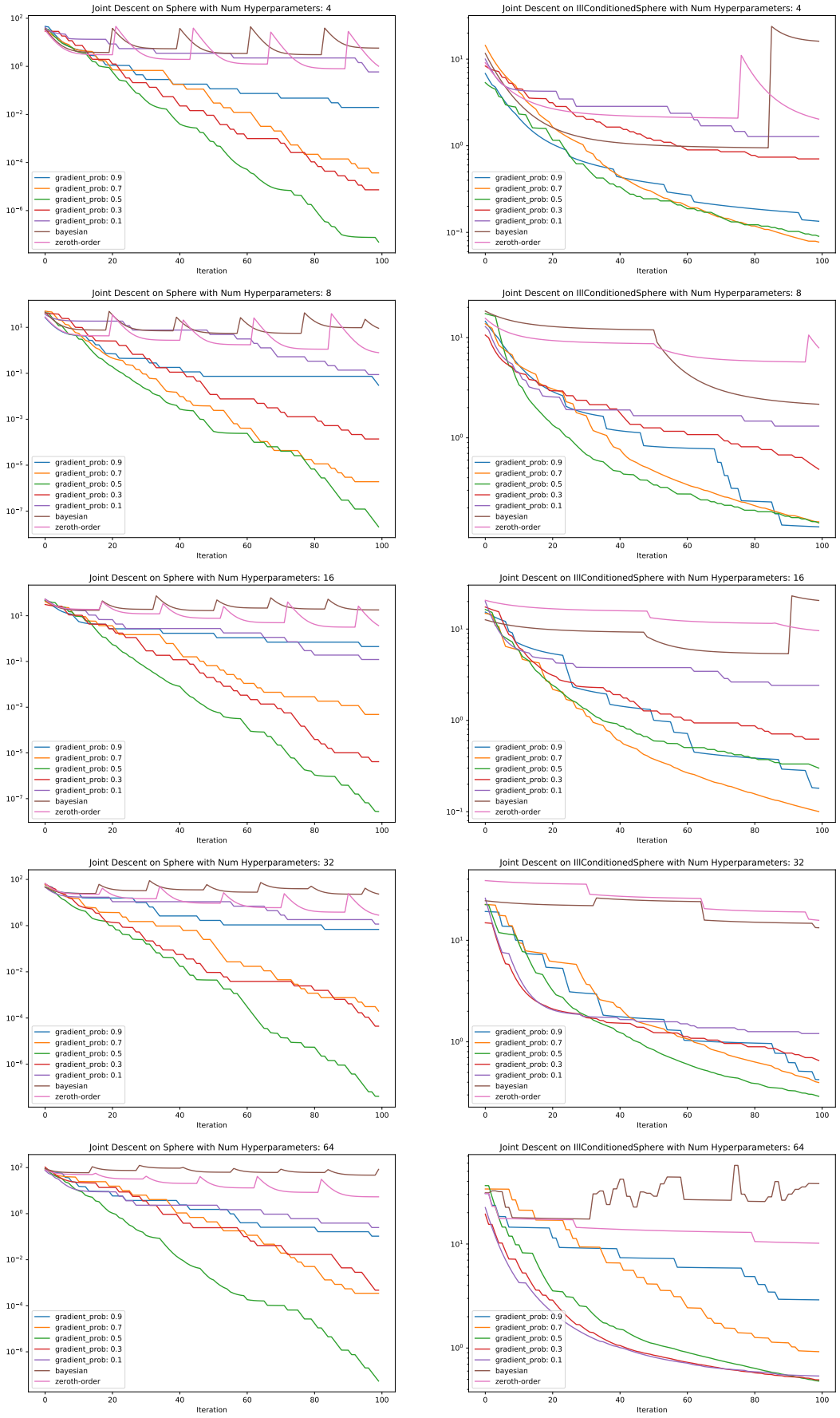


Figure 5: Convergence plot for Joint Descent on the Sphere and Ill-Conditioned Sphere with number of training variables set to 30 and varying number of hyperparameters.

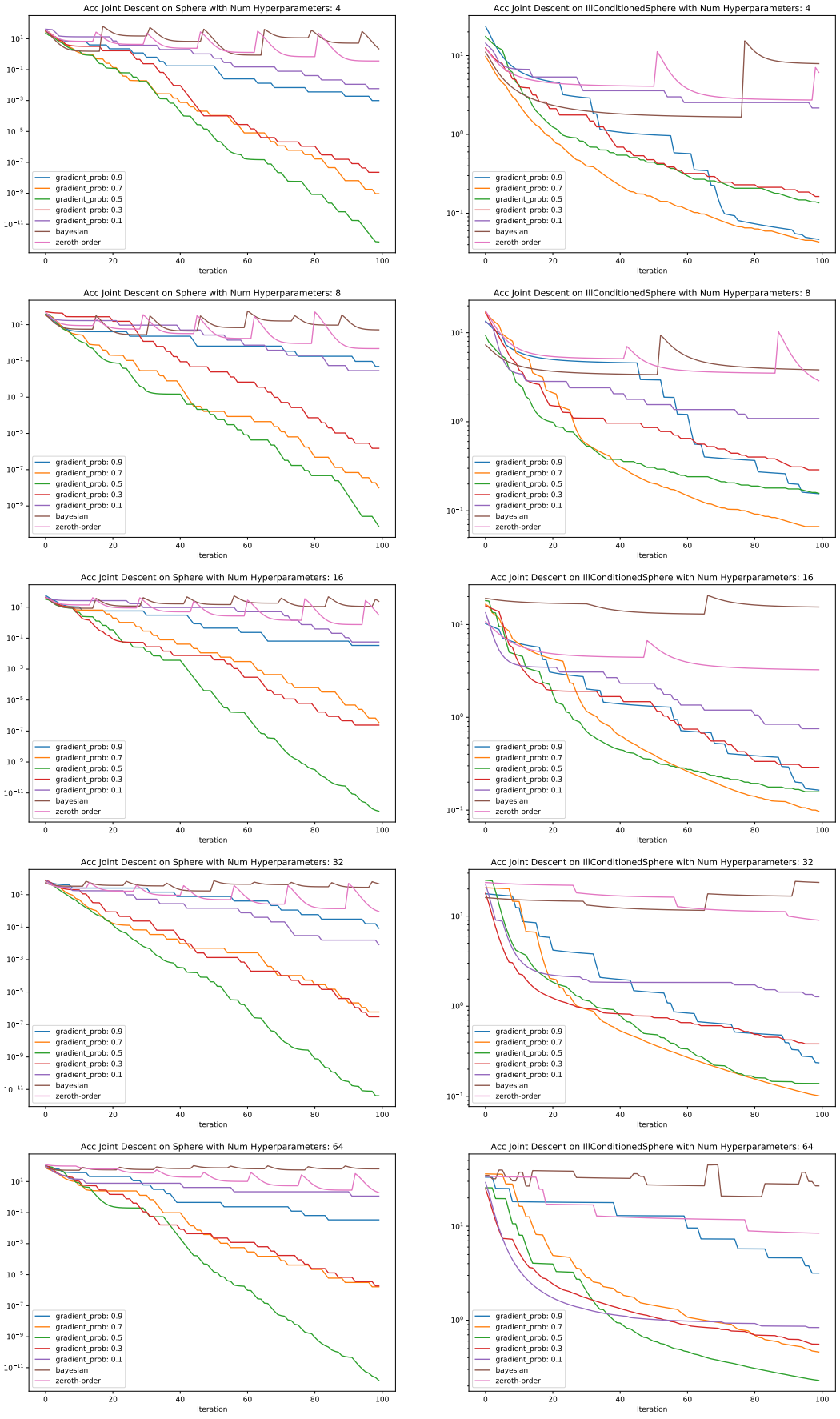


Figure 6: Convergence plot for Accelerated Joint Descent on the Sphere and Ill-Conditioned Sphere with number of training variables set to 30 and varying number of hyperparameters.