

Figure 3: Reconstruction loss on noisy MNIST samples. Error bars (negligibly small) show one standard deviation (in log-scale) in average loss ( $B=2$ , each containing 30 batched measurements), over 2 random seeds and 2 random initializations per seed.

## A APPENDIX

Subsection A.1 details additional experiments. Subsection A.2 and Subsection A.3 include complete proofs of the results Theorem 5.2 and Theorem 5.3 whose outlines were included in the main body.

### A.1 ADDITIONAL EXPERIMENTS

In this section, we first describe details of the experiments (Subsubsection A.1.1) before including supplemental experiments (Subsubsection A.1.2 and Subsubsection A.1.3) that extend those in Section 6. In Subsubsection A.1.2, we perform a denoising task on the original MNIST dataset itself. In Subsubsection A.1.3, we include additional plots from the transfer learning setup with MNIST digits. Again, both experiments demonstrate the practical value of incorporating a generative coefficient prior  $G$  even when it is not explicitly part of the data-generating process.

#### A.1.1 EXPERIMENTAL DETAILS

All experiments were run on a NVIDIA Tesla V100SXM2 GPU from an internal cluster, and each took no longer than 2 hours. Constant optimization parameters (e.g. learning rates) were chosen by inspection on initial experiments such that losses converged, and used throughout thereafter.  $n$ ,  $m$ , and  $k$  parameters were determined largely by the choice of toy problems. The implementation of kSVD is available at <https://github.com/nel215/ksvd> under the Apache License. For the pretrained MNIST VAE, the decoder architecture is  $S(W_2 \text{ReLU}(W_1 z))$ , where  $S(\cdot)$  is the sigmoid function,  $W_1 \in \mathbb{R}^{400 \times 20}$ , and  $W_2 \in \mathbb{R}^{784 \times 400}$ . In the case of the transfer learning on MNIST setup, we perform a blur via `ImageFilter.BoxBlur` with radius 2, a colorshift via `torchvision.transforms.ColorJitter` with hue set to 0.4 and brightness, contrast, and saturation set to 1, or a horizontal flip via `torchvision.transforms.RandomHorizontalFlip` with probability 1. The MNIST VAE model and code were provided by a collaborator; acknowledgement and training code to be added following anonymous submission.

#### A.1.2 MNIST DENOISING

In this section, we apply the generative prior  $G$  to denoise *real* MNIST images. In other words, we draw noisy samples  $y = y' + \eta$ , where  $y'$  is drawn from a training set of MNIST images and  $\eta \sim N(0, \sigma^2 I)$  with  $\sigma = 0.05$ , and try to find  $A^*$  such that  $y \sim A^* G(\cdot)$ . As before, autodiff and altmin are more effective at denoising than traditional dictionary learning methods. This again validates the choice of model for the more realistic setting (corresponding to large, non-Gaussian noise) in which the generative coefficient prior is chosen based on a related task, but its composition with a linear transformation likely does not fully describe the data. Methods are trained using only the training set, and their reconstruction error is evaluated using only the test set. Results are shown in Figures 3 and 4. As before, latent variables are rounded to 4-digit precision.

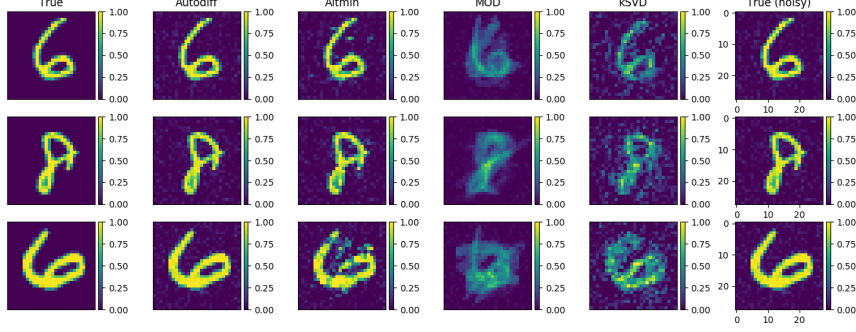


Figure 4: Example reconstructions of MNIST by each method; each row is a different sample. The leftmost column shows the noiseless ground truth, and the rightmost column shows the noisy measurements provided as input to each method.

In addition, all experiments thus far have used only the top  $k$  coefficients, each rounded to 4 digits, to reconstruct the desired signal for the MOD and kSVD baselines. The motivation is that each reconstruction method being compared is permitted to use  $k$  free parameters to reconstruct each image, analogously to compressing images in the dataset via  $k$  bits. Since the only free parameter for autodiff and altmin,  $z$ , is  $k$ -dimensional, this suggests that a fair comparison with baselines should only allow for  $k$ -parameter descriptors as well. However, in Figures 5a and 5b we confirm that autodiff and altmin still provide much better reconstructions than MOD or kSVD without this rounding or truncation, even when allowed to use  $n > k$  free parameters (coefficients).

#### A.1.3 TRANSFORMED MNIST DENOISING

The analogous figure to Figure 2 for the flipped dataset is included here in Figure 6b

#### A.2 PROOFS AND DETAILS FOR THEOREM 5.2

Let’s start out with some basic manipulations that will come in handy. Note that these assumptions hold at the start of the update step, as they draw upon the assumed behavior of the decode step.

**Lemma A.1** (Closeness between  $y$  and  $AG(z)$ .) *Let  $A$  be an arbitrary matrix in  $B_2(A^*, \Delta)$ ,  $\Sigma = A - A^*$ , and  $y = A^*G(z^*) + \eta$  for some unknown  $z^*$ . Since  $A \in B_2(A^*, \Delta)$ , we can apply the optimization oracle to  $y$  with our current guess  $A$  to give an estimate  $z$ . The reconstruction error of estimates  $A$  and  $z$  in explaining the measurement  $y$  is then bounded as follows:*

$$\|y - AG(z)\| \leq \|A^* - A\| \cdot R + \theta + \|\eta\| = \|\Sigma\|R + \theta + \|\eta\|$$

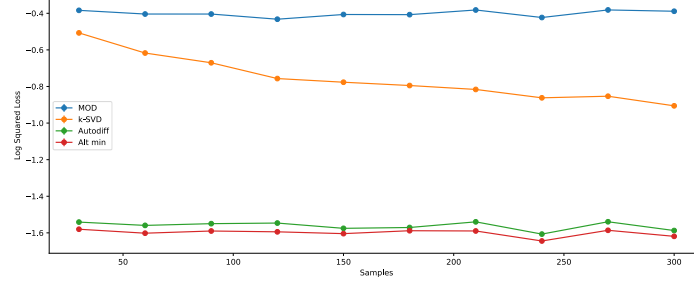
By assumption on the optimization oracle,  $\|y - AG(z)\| \leq \min_{\|\tilde{z}\| \leq \frac{R}{L}} \|y - AG(\tilde{z})\| + \theta$ . We can express the first term as a function of  $\|A - A^*\| \|\Sigma\|$  as follows:

$$\min_{\|\tilde{z}\| \leq \frac{R}{L}} \|y - AG(\tilde{z})\| \leq \|y - AG(z^*)\| \leq \underbrace{\|y - A^*G(z^*)\|}_{=\|\eta\|} + \underbrace{\|(A^* - A)G(z^*)\|}_{\leq \|A^* - A\| \cdot R}$$

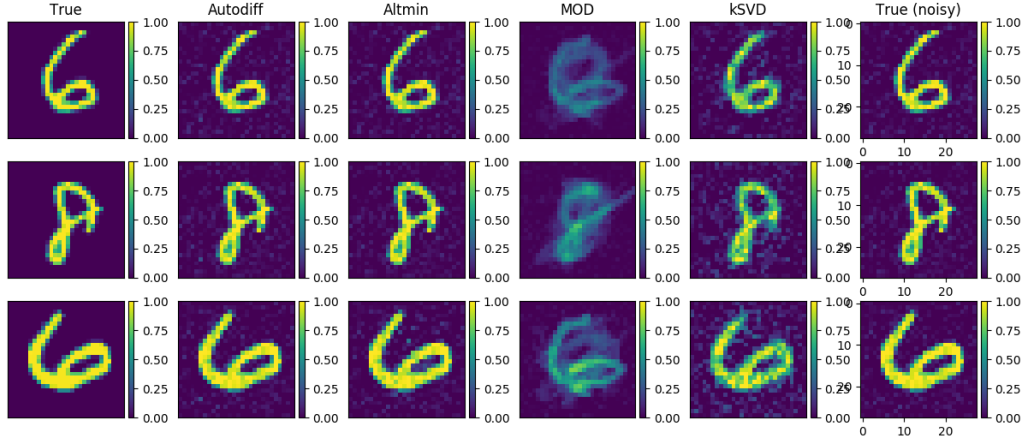
From this, the result follows. Note also that, by initialization and the projection step, we have that  $\|\Sigma\| \leq n\Delta$ .

**Lemma A.2** (Closeness between  $G(z)$  and  $G(z^*)$ .) *Let  $A$  be an arbitrary matrix and  $\Sigma = A - A^*$ . Let  $z$  be the output of the optimization oracle corresponding to  $A$  with any fixed measurement generated as  $y = A^*G(z^*) + \eta$ . Then we have:*

$$\|G(z) - G(z^*)\| \leq \frac{2\|\Sigma\|R + \theta + \delta + \|\eta\|}{\gamma}$$

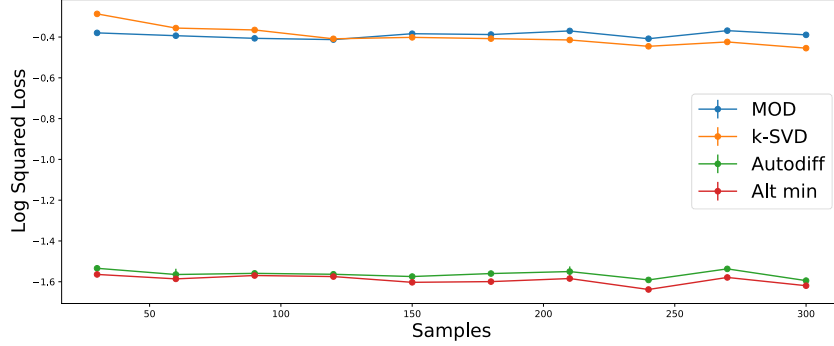


(a) Reconstruction loss on noisy MNIST samples, where baseline reconstructions use all  $n$  (not just  $k$ ) coefficients. Error bars (negligibly small) show one standard deviation (in log-scale) in average loss ( $B=2$ , each containing 30 batched measurements), over 2 random seeds and 2 random initializations per seed.

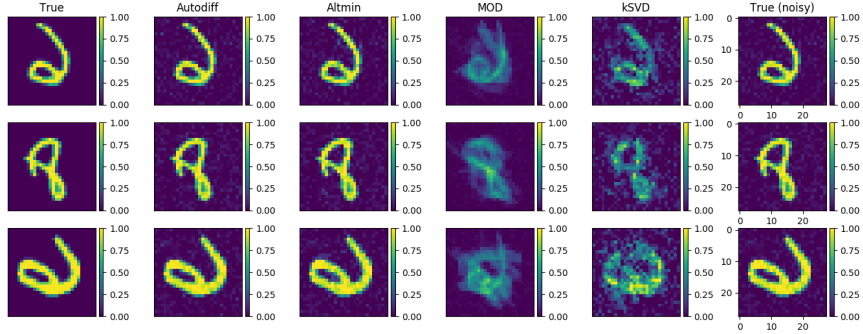


(b) Example reconstructions of MNIST, where baselines use all  $n > k$  coefficients, by each method; each row is a different sample. The leftmost column shows the noiseless ground truth, and the rightmost column shows the noisy measurements provided as input to each method.

Figure 5



(a) Error bars show one standard deviation (in log-scale) over 4 trials, each with  $n_B = 2$  and  $B = 30$ .



(b) Example reconstructions on flipped dataset by each method; each row is a different sample. The leftmost column shows the noiseless ground truth.

$$\gamma \|G(z) - G(z^*)\| - \delta \leq \|A^*G(z) - A^*G(z^*)\| \text{ by S-REC} \quad (1)$$

$$\leq \|AG(z) - A^*G(z^*)\| + \|A^*G(z) - AG(z)\| \quad (2)$$

$$\begin{aligned} & \underbrace{\|AG(z) - A^*G(z^*)\|}_{\leq \|\Sigma\|R + \theta + \|\eta\| \text{ by A.1}} + \underbrace{\|A^*G(z) - AG(z)\|}_{\leq \|\Sigma\|R} \\ & \leq 2\|\Sigma\|R + \theta + \|\eta\| \end{aligned} \quad (3)$$

Rearranging terms finishes the proof.

**Remark 3** [Theorem A.2](#) implies that, given a good estimate  $A$  of  $A^*$ , the optimization oracle can find a good estimate  $z$  of  $z^*$ .

**Lemma A.3** (Closeness between  $A\mathbb{E}[G(z^*)G(z^*)^T]C^{-1}$  and  $A$ , for arbitrary  $A$ .)

$$\|A\mathbb{E}[G(z^*)G(z^*)^T]C^{-1} - A\| \leq \|A\| \cdot \|C^*\| \cdot \nu$$

$$\begin{aligned} \|A\mathbb{E}[G(z^*)G(z^*)^T]C^{-1} - A\| &= \|A\mathbb{E}[G(z^*)G(z^*)^T]C^{-1} - A\mathbb{E}[G(z^*)G(z^*)^T](C^*)^{-1}\| \\ &\leq \|A\| \cdot \|\mathbb{E}[G(z^*)G(z^*)^T]\| \cdot \nu \\ &= \|A\| \cdot \|C^*\| \cdot \nu \end{aligned}$$

**Lemma A.4** (Non-centered vector Bernstein.) Suppose  $\vec{x}^1, \dots, \vec{x}^P$  are i.i.d. vector-valued random variables such that, for all  $i$ ,  $\mathbb{E}[\vec{x}^i] = \vec{m}$ ,  $\|\vec{x}^i\| \leq R$ , and  $\mathbb{E}[\|\vec{x}^i\|^2] \leq \sigma^2$ . Then for  $0 < \epsilon < \frac{\sigma^2}{R}$ ,

$$P\left(\left\|\left(\frac{1}{P} \sum_{i=1}^P \vec{x}^i\right) - \mathbb{E}[\vec{x}^i]\right\| \geq \epsilon\right) \leq \exp\left(\frac{-P\epsilon^2}{16\sigma^2} + \frac{1}{4}\right)$$

We will use the vector Bernstein inequality from Lemma 18 of (Kohler & Lucchi, 2017). Let  $\bar{w}^i = \bar{x}^i - \mathbb{E}[\bar{x}^i]$ . Then note that  $\mathbb{E}[\bar{w}^i] = 0$  and  $\|\bar{w}^i\| \leq \|\bar{x}^i\| + \|\mathbb{E}[\bar{x}^i]\| = 2R$ . Additionally, we have:

$$\begin{aligned}\mathbb{E}[\|\bar{w}^i\|^2] &= \mathbb{E}[(\bar{x}^i - \mathbb{E}[\bar{x}^i])^T (\bar{x}^i - \mathbb{E}[\bar{x}^i])] \\ &= \mathbb{E}[\bar{x}^{iT} \bar{x}^i - 2\bar{x}^{iT} \mathbb{E}[\bar{x}^i] + \mathbb{E}[\bar{x}^i]^T \mathbb{E}[\bar{x}^i]] \\ &= \mathbb{E}[\|\bar{x}^i\|^2] - 2\|\mathbb{E}[\bar{x}^i]\|^2 + \|\mathbb{E}[\bar{x}^i]\|^2 \\ &= \mathbb{E}[\|\bar{x}^i\|^2] - \|\mathbb{E}[\bar{x}^i]\|^2 \\ &\leq \mathbb{E}[\|\bar{x}^i\|^2] + \|\mathbb{E}[\bar{x}^i]\|^2 \\ &\leq 2\mathbb{E}[\|\bar{x}^i\|^2] \\ &\leq 2\sigma^2\end{aligned}$$

We then apply the cited vector Bernstein inequality to  $\bar{w}^i$  and obtain the stated result. For the rest of the proof, we will rely on certain definitions and results from (Arora et al., 2015), included here for completeness. We begin with a careful notion of vector correlation:  $((\alpha, \beta, \epsilon)$ -correlated) Consider an iterative algorithm with steps of the form  $v^{s+1} = v^s - \eta g^s$  and desired solution  $v^*$ . We say the vector  $g^s$  is  $(\alpha, \beta, \epsilon_s)$ -correlated with  $v^s - v^*$  if

$$\langle g, v^s - v^* \rangle \geq \alpha \|v^s - v^*\|^2 + \beta \|g^s\|^2 - \epsilon_s$$

In our case, we will think of  $v^s$  as a column of iterate  $A^s$ ,  $v^*$  as the corresponding column of  $A^*$ , and  $g^s$  and  $\hat{g}^s$  as they are already defined in Algorithm 1

**Lemma A.5** ( $g^s$  is correlated with  $A^s - A^*$ ) Let  $\zeta = 10n \frac{\theta R}{\gamma}$ . Then we have column-wise correlation between matrices  $g^s$  and  $A^s - A^*$ :  $(g^s)_i$  is  $(\frac{1}{4}, \frac{1}{25}, 4\zeta^2)$ -correlated with  $(A^s)_i - (A^*)_i$ .

For convenience, we write  $A$  in place of  $A^s$ . Also, note that since  $\eta$  is entrywise iid and zero-mean,  $\mathbb{E}_{z, \eta}[\eta G(z)^T] = \mathbf{0}$ . Then:

$$g^s = \mathbb{E}[(AG(z) - A^*G(z^*))G(z)^T]C^{-1} \quad (4)$$

$$= A\mathbb{E}[G(z)G(z)^T]C^{-1} - A^*\mathbb{E}[G(z^*)G(z)^T]C^{-1} \quad (5)$$

Consider  $\mathbb{E}[G(z^*)G(z)^T]C^{-1}$  alone to start.

$$\mathbb{E}[G(z^*)G(z)^T]C^{-1} = \mathbb{E}[G(z^*)G(z^*)^T]C^{-1} + \mathbb{E}[G(z^*)(G(z) - G(z^*))^T]C^{-1}$$

By assumption,  $\mathbb{E}[G(z^*)G(z^*)^T]C^{-1}$  is close to the identity, whereas  $\mathbb{E}[G(z^*)(G(z) - G(z^*))^T]C^{-1}$  is purely an error term we seek to bound as follows.

$$\begin{aligned}\|\mathbb{E}[G(z^*)(G(z) - G(z^*))^T]C^{-1}\|^2 &\leq \mathbb{E}[\|G(z^*)(G(z) - G(z^*))^T\|^2] \cdot \|C^{-1}\|^2 \\ &\leq \mathbb{E}[\|G(z^*)\|^2 \cdot \|G(z) - G(z^*)\|^2] \cdot \|C^{-1}\|^2 \\ &\leq R^2 \left( \frac{2\|\Sigma\|R + \theta + \delta + \|\eta\|}{\gamma} \right)^2 \cdot \|C^{-1}\|^2 \text{ by Lemma A.2}\end{aligned}$$

Thus,

$$\|\mathbb{E}[G(z^*)(G(z) - G(z^*))^T]C^{-1}\| \leq \frac{2\|\Sigma\|R^2 + (\theta + \delta + \|\eta\|)R}{\gamma} \cdot \|C^{-1}\|$$

Now, consider

$$\mathbb{E}[G(z)G(z)^T] = \mathbb{E}[G(z^*)G(z^*)^T] + \mathbb{E}[(G(z) - G(z^*))G(z)^T] + \mathbb{E}[G(z^*)(G(z) - G(z^*))^T]$$

Once again, we bound the “error terms”:

$$\begin{aligned}\|\mathbb{E}[(G(z) - G(z^*))G(z)^T]C^{-1} + \mathbb{E}[G(z^*)(G(z) - G(z^*))^T]C^{-1}\| &\leq \\ &\frac{4\|\Sigma\|R^2 + 2(\theta + \delta + \|\eta\|)R}{\gamma} \cdot \|C^{-1}\|\end{aligned}$$

We used the same bound as before for *both* terms because, by assumption on the optimization, we still have  $\|G(z)\| \leq R$ . Let's return to the original expression.

$$\begin{aligned} g^s &= A\mathbb{E}[G(z)G(z)^T]C^{-1} - A^*\mathbb{E}[G(z^*)G(z)^T]C^{-1} \\ &= A\mathbb{E}[G(z^*)G(z^*)^T]C^{-1} + AE_1 - A^*\mathbb{E}[G(z^*)G(z^*)^T]C^{-1} + A^*E_2 \\ &= (A - A^*)\mathbb{E}[G(z^*)G(z^*)^T]C^{-1} + AE_1 + A^*E_2 \\ &= (A - A^*) + AE_1 + A^*E_2 + E_3 \end{aligned}$$

By Lemma A.3,  $\|E_3\| \leq \nu\|C^*\| \cdot \|A - A^*\|$ . By the previous computations,  $\|E_2\| \leq 2\frac{\|\Sigma\|R^2 + (\theta + \delta + \|\eta\|)R}{\gamma}\|C^{-1}\|$  and  $\|E_1\| \leq 4\frac{\|\Sigma\|R^2 + (\theta + \delta + \|\eta\|)R}{\gamma}\|C^{-1}\|$ . By assumption that the true dictionary has unit columns,  $\|A^*\| \leq n$ , and by initialization and projections,  $\|A - A^*\| \leq n\Delta$ , such that  $\|A\| \leq n(1 + \Delta)$ . Combining all of these, we have

$$\begin{aligned} \|AE_1 + A^*E_2 + E_3\| &\leq \nu\|C^*\| \cdot \|\Sigma\| + n(1 + \Delta)\|C^{-1}\| \frac{4\|\Sigma\|R^2 + 2(\theta + \delta + \|\eta\|)R}{\gamma} \\ &\quad + n\|C^{-1}\| \frac{2\|\Sigma\|R^2 + (\theta + \delta)R}{\gamma} \\ &= \|\Sigma\| \left( \nu\|C^*\| + 4n\Delta \frac{R^2}{\gamma}\|C^{-1}\| + 6n \frac{R^2}{\gamma}\|C^{-1}\| \right) \\ &\quad + 2n(1 + \Delta)\|C^{-1}\| \frac{(\theta + \delta + \|\eta\|)R}{\gamma} + n\|C^{-1}\| \frac{(\theta + \delta + \|\eta\|)R}{\gamma} \\ &\leq \|\Sigma\| \left( \nu\|C^*\| + 10n\|C^{-1}\| \frac{R^2}{\gamma} \right) + 5n\|C^{-1}\| \frac{(\theta + \delta + \|\eta\|)R}{\gamma} \end{aligned}$$

Without yet specifying exactly the required order  $\Delta$  of the initialization, we have assumed that  $\Delta < 1$  to simplify the expression in the final step. We now use the assumption governing the interaction between these variables:  $\left( \nu\|C^*\| + 10n\|C^{-1}\| \frac{R^2}{\gamma} \right) \leq \frac{1}{4}$ . Call the second term  $\zeta = 5n\|C^{-1}\| \frac{(\theta + \delta + \|\eta\|)R}{\gamma}$ . Then we apply Lemma 15 of (Arora et al., 2015) with  $\alpha = \frac{1}{4}$ , making use of our interplay of parameters assumption. Noting that all of the matrix bounds above suffice as column-wise upper bounds too and applying Lemma 15, we have that  $g_i^s$ , the  $i^{th}$  column of  $g^s$ , is  $(\frac{1}{4}, \frac{1}{25}, 4\zeta^2)$ -correlated with  $A_i^s - A_i^*$ .

In practice, we don't have access to  $g^s$  exactly; instead, we approximate the expectation with  $P$  fresh i.i.d. samples,  $\hat{g}^s = \frac{1}{P} \sum_{p=0}^{P-1} (AG(z^p) - y^p)G(z^p)^T C^{-1}$ . If  $\hat{g}^s$  is close enough to  $g^s$ , however, then  $\hat{g}^s$  will remain correlated-w.h.p. with the correct direction  $A^s - A^*$ :

We quantify this closeness in the following lemma.

**Lemma A.6** ( $\hat{g}^s$  is close to  $g^s$ ) For  $0 < \epsilon < (n\Delta R^2 + \theta R)$  and  $\mu = n\Delta R^2\|C^{-1}\| + (\theta + N)R\|C^{-1}\|$ ,

$$P\left(\left\|\hat{g}_i^s - g_i^s\right\| \geq \epsilon\right) \leq \exp\left(\frac{-P\epsilon^2}{16\mu^2} + \frac{1}{4}\right)$$

We will apply Lemma A.4 with  $\bar{x}^p = \left((AG(z^p) - y^p)G(z^p)^T C^{-1}\right)_i$

$$\left\|\left((AG(z^p) - y^p)G(z^p)^T C^{-1}\right)_i\right\| \leq \|(AG(z^p) - y^p)G(z^p)^T C^{-1}\| \quad (6)$$

$$\leq (\|\Sigma\|R + \theta + N)R\|C^{-1}\| \text{ by Lemma A.1} \quad (7)$$

$$\leq n\Delta R^2\|C^{-1}\| + (\theta + N)R\|C^{-1}\| = \mu \quad (8)$$

Applying Lemma A.4 with  $\|\bar{x}^i\| \leq \mu$ ,  $\|\bar{x}^i\|^2 \leq \mu^2$ , we have that for  $0 < \epsilon < \mu$ ,

$$P\left(\left\|\left(\frac{1}{P} \sum_{i=1}^P \bar{x}^i\right) - \mathbb{E}[\bar{x}^i]\right\| \geq \epsilon\right) \leq \exp\left(\frac{-P\epsilon^2}{16\mu^2} + \frac{1}{4}\right) \quad (9)$$

$$\rightarrow P\left(\left\|\hat{g}_i^s - g_i^s\right\| \geq \epsilon\right) \leq \exp\left(\frac{-P\epsilon^2}{16\mu^2} + \frac{1}{4}\right) \quad (10)$$

Finally, we require the probabilistic version of correlation introduced by Arora et al. (2015):  $((\alpha, \beta, \epsilon)$ -correlated-w.h.p.) Again consider an iterative algorithm with steps of the form  $v^{s+1} = v^s - \eta g^s$  and desired solution  $v^*$ . Now, assume  $g^s$  is a random vector. We say the random vector  $g^s$  is  $(\alpha, \beta, \epsilon_s)$ -correlated-w.h.p. with  $v^s - v^*$  if with probability at least  $1 - n^{-\omega(1)}$

$$\langle g, v^s - v^* \rangle \geq \alpha \|z^s - z^*\|^2 + \beta \|g^s\|^2 - \epsilon_s$$

We now apply Equation A.2 to show that  $\hat{g}^s$  is correlated-w.h.p. with  $A^s - A^*$ .

**Theorem A.7** ( $\hat{g}^s$  is correlated-w.h.p. with  $A^s - A^*$ ) Draw  $P = n^{2L} \log^2 n$  fresh, i.i.d. samples at every iteration, and assume initialization  $A^0 \in B_2(A^*, \Delta)$ . Then  $\hat{g}_i^s$  is  $(\frac{1}{4}, \frac{1}{25}, 100n^2 \|C^{-1}\|^2 \frac{(\theta + \delta + \|\eta\|)^2 R^2}{\gamma^2} + \frac{(n\Delta R + \theta + N)R \|C^{-1}\|}{n^L})$ -correlated-w.h.p. with  $A_i^s - A_i^*$ .

Recall that  $\zeta = 5n \|C^{-1}\| \frac{(\theta + \delta + \|\eta\|)R}{\gamma}$  and  $\mu = n\Delta R^2 \|C^{-1}\| + (\theta + N)R \|C^{-1}\|$ . For a particular setting of  $\epsilon$  and number of samples per iteration  $P$  plugged into Lemma A.6 we will obtain that

$\hat{g}_i^s$  is  $(\frac{1}{4}, \frac{1}{25}, 4\zeta^2 + \epsilon)$ -correlated with  $A_i^s - A_i^*$  with probability  $1 - \exp\left(\frac{-P\epsilon^2}{16\mu^2} + \frac{1}{4}\right)$ . To be

“correlated-w.h.p.” in the language of Arora et al. (2015), we require  $n^{-\omega(1)} \geq \exp\left(\frac{-P\epsilon^2}{16\mu^2} + \frac{1}{4}\right)$ .

One way to achieve this is by setting  $\epsilon = \frac{(n\Delta R + \theta + N)R \|C^{-1}\|}{n^L} = \frac{\mu}{n^L}$  and  $P = n^{2L} \log^2 n$ . Then  $\frac{-P\epsilon^2}{16\mu^2} = \frac{-n^{2L} \log^2 n}{n^{2L}} = \frac{-\log^2 n}{16} \leq \log(n^{-\omega(1)})$  as desired, since  $P$  satisfies  $P \geq \omega(1) \log n$ .

**Theorem A.8** (Geometric optimization convergence.) With the conditions of Theorem A.7 and with step size  $\eta = \frac{2}{25}$ , alternating minimization converges geometrically, but with a bias:

$$\begin{aligned} \mathbb{E}[\|A_i^t - A_i^*\|^2] &\leq (1 - \frac{1}{25})^t \|A_i^0 - A_i^*\|^2 + \\ &\quad \left( \frac{400n^2 \|C^{-1}\|^2 (\theta + \delta + N)^2 R^2}{\gamma^2} + \frac{4(n\Delta R + \theta + N)R \|C^{-1}\|}{n^L} \right) \end{aligned}$$

This theorem follows as a direct application of Theorem 40 of Arora et al. (2015) to our Theorem A.7.

Unfortunately, this bias term is non-negligible; even if  $A^*$  is invertible s.t.  $\delta = 0$ , we can only reduce by assuming a more powerful optimization oracle, which decreases  $\theta$ . It is an open problem to find a different update rule, or a refined analysis, with lower bias analogously to e.g. Appendix B.2 of Arora et al. (2015).

### A.3 PROOFS AND DETAILS FOR THEOREM 5.3

The full terms of the Weight Distribution Condition of Hand & Voroninski (2018) are as follows: (Weight Distribution Condition) A matrix  $W \in \mathbb{R}^{n \times k}$  with  $i^{th}$  row  $w_i$  satisfies the *Weight Distribution Condition* (WDC) if  $\forall x, y \neq 0 \in \mathbb{R}^k$ ,

$$\left\| \sum_{i=1}^n 1_{w_i x > 0} 1_{w_i y > 0} \cdot w_i w_i^T - Q_{x,y} \right\| \leq \epsilon, \text{ with } Q_{x,y} = \frac{\pi - \theta}{2\pi} I_k + \frac{\sin \theta}{2\pi} M_{\frac{x}{\|x\|} \leftrightarrow \frac{y}{\|y\|}}$$

where  $M_{a \leftrightarrow b} \in \mathbb{R}^{k \times k}$  is the matrix such that  $Ma = b$ ,  $Mb = a$ , and  $Mz = 0 \forall z \in \text{span}(x, y)^\perp$ . Recall that we wish to learn  $A^*$ , and at each iteration we minimize a modified loss function  $f =$

$\|AG(z) - A^*G(z^*)\|^2$  where  $A \approx A^*$ . We will show that the optimization landscape of  $f$  closely resembles that of  $\tilde{f}(z) = \|AG(z) - AG(z^*)\|^2$  when  $A$  is a good approximation of  $A^*$ , and use this to prove [Theorem 5.3](#). We begin with the following crucial lemma, as described:

**Lemma A.9** (*A close to  $A^*$  satisfies the RRIC*) *If  $A^*$  satisfies the RRIC with respect to  $G$  with constant  $\epsilon$ , then any  $A$  sufficiently close to  $A^*$  satisfies the RRIC with constant  $c\epsilon + 2\|A^* - A\| \cdot \|A^*\| + \|A^* - A\|^2$*

$$\begin{aligned} & |\langle A(G(x_1) - G(x_2)), A(G(x_3) - G(x_4)) \rangle - \langle A^*(G(x_1) - G(x_2)), A^*(G(x_3) - G(x_4)) \rangle| \leq \\ & \quad |\langle (A^* - A)(G(x_1) - G(x_2)), A^*(G(x_3) - G(x_4)) \rangle| + \\ & \quad |\langle A^*(G(x_1) - G(x_2)), (A^* - A)(G(x_3) - G(x_4)) \rangle| + \\ & \quad |\langle (A^* - A)(G(x_1) - G(x_2)), (A^* - A)(G(x_3) - G(x_4)) \rangle| \leq \\ & \quad (2\|A^* - A\| \cdot \|A^*\| + \|A^* - A\|^2) \cdot \|(G(x_1) - G(x_2))\| \cdot \|(G(x_3) - G(x_4))\| \end{aligned}$$

Then, in order to reduce to the setting of [\(Hand & Voroninski, 2018\)](#), we will argue that the gradients of  $\|AG(x) - A^*G(x^*)\|^2$  are close to the gradients of  $\|AG(x) - AG(x^*)\|^2$ . We now know that  $A$  satisfies the RRIC with respect to  $G$  with constant  $c$ . Furthermore, their original proof applied immediately to objective function  $\|AG(x) - AG(x^*)\|^2$  relies on making approximations to the objective function's gradient. We will show that the order of error in the approximation to the gradient is unchanged when adding an additional error term, which comes from using the gradient of the objective function  $\|AG(x) - A^*G(x^*)\|^2$  instead of that of  $\|AG(x) - AG(x^*)\|^2$ . Thus, the original proof will go through. We also know that  $G$  satisfies the WDC with constant  $c > \epsilon$  and proceed by substituting  $\epsilon$  in their paper with  $c$ .

Observe that if  $G(z) = (\prod_{i=d}^1 W_{i,+} z)$ , then the gradient of  $\frac{1}{2}\|AG(z) - A^*G(z^*)\|^2$  with respect to  $z$  (letting  $W = \prod_{i=d}^1 W_{i,+}$ ) is

$$W^T A^T A W z - W^T A^T A^* W z^*$$

This expression is “close” to the gradient of  $\|AG(z) - AG(z^*)\|^2$  with respect to  $z$ ,  $W^T A^T A W z - W^T A^T A W z^*$  (i.e. where the  $A^*$  is replaced with  $A$ ), with error term

$$\|W^T A^T (A - A^*) W z^*\| \leq \|W\|^2 \|A\| \cdot \|A - A^*\| \cdot \|z^*\|$$

By assumption that  $G$  satisfies the WDC with constant  $c$ ,  $\|W\|^2 \leq (\frac{1}{2} + c)^d$ . If  $\|A - A^*\| \leq \frac{c}{\|A\|}$ , then equation (26) in the Hand/Voroninski paper holds with constant 3 instead of 2. The effect of this change in constant percolates down to equation (30), but merely increases the universal constant  $\tilde{K}$ . The rest of the proof then holds identically. The final statement is that, if  $G$  and  $A^*$  each satisfy the WCD and RRIC respectively with constants  $\epsilon$ , and if

$$\|A - A^*\| \leq \frac{c}{\|A\|} = \frac{\epsilon + 2\|A^* - A\| \cdot \|A^*\| + \|A^* - A\|^2}{\|A\|}$$

then Theorem 4 of [\(Hand & Voroninski, 2018\)](#) holds with  $\epsilon' = c$  instead of the original  $\epsilon$ . This yields [Theorem 5.3](#) as desired.