Supplementary Material

S1. Adaptive Decoding

Limited training data in practical settings can limit the inferential accuracy of the learned autoencoder model, and we may have $x_0 \neq D(E(x_0))$ for a given to-be-revised sequence x_0 (particularly if $p_X(x_0)$ is low). In this case, even when $z^* = E(x_0)$ solves our latent-factor optimization, our REVISE procedure can return a different sequence than x_0 (despite not expecting any associated outcome-improvement).

To ensure that our methods simply return the initial x_0 when no superior revision can be identified, we replace our decoder model $p_D(x \mid z)$ with an adaptive variant $p_{D_{x_0}}(x \mid z)$ that is efficiently defined once $x_0 = \left(s_1^{(x_0)}, \ldots, s_{T_{x_0}}^{(x_0)}\right)$ is specified at test time. Like before, we write $D_{x_0}(z)$ to denote the (beam-search approximated) most-likely decoding with respect to $p_{D_{x_0}}$. Recall from our definition in (4), π_t is the vector of symbol-probabilities output by our decoder RNN $\mathcal{D}$ to compute p_D. Using the indexing notation $\pi_t[s_t]$ to denote the decoder RNN's approximation of $p(s_t, \mid s_1 \ldots, s_{t-1})$, we let $\pi_t^{(x_0)}$ denote particular conditional-probability values output by $\mathcal{D}$ when the initial hidden state is $z = E(x_0)$.

For any $x = (s_1, \ldots, s_T) \in \mathcal{X}$, we define:

$$p_{D_{x_0}}(x \mid z) = \prod_{t=1}^{T} \widetilde{\pi}_t[s_t] \quad \text{where for } t = 1, \ldots, T, \ s \in \mathcal{S}: \quad \widetilde{\pi}_t[s] = \begin{cases} \pi_t[s] + \beta_t^{(x_0)} & \text{if } s = s_t^{(x_0)} \\ \pi_t[s] - \frac{1}{|\mathcal{S}|}\beta_t^{(x_0)} & \text{otherwise} \end{cases} \tag{15}$$

$$\text{and } \beta_t^{(x_0)} = \max_{s \in \mathcal{S}} \pi_t^{(x_0)}[s] - \pi_t^{(x_0)}[s_t^{(x_0)}] \geqslant 0 \quad \text{for } t = 1, \ldots, T_{x_0}$$

At each time step, the $\beta_t^{(x_0)}$ measure any probability gap between the most likely symbol under p_D and the actual sequence x_0 when our decoder model $\mathcal{D}$ is applied to $E(x_0)$. Thus, the definition in (15) ensures $D_{x_0}(E(x_0)) = x_0$. When revising sequences using this adaptive decoding procedure, we compute all $\beta_t^{(x_0)}$ by first decoding from $E(x_0)$ before beginning the latent z-optimization in the REVISE procedure. These values are stored so that we can subsequently decode from the optimal latent-configuration z^* with respect to $p_{D_{x_0}}$ rather than p_D.

According to our adaptive decoding definition, x_0 is more likely than any other sequence under $p_{D_{x_0}}(x \mid E(x_0))$, and $p_{D_{x_0}}$ is very easy to derive from p_D (no additional model besides our original $\mathcal{D}$ is needed). Furthermore, the (beam-search) maximizer of $p_{D_{x_0}}$ can be used to decode from any latent z values, resulting in a mapping that is slightly more biased toward x_0 than decoding with respect to p_D. Finally, we note that if x^* is produced by D_{x_0} rather than D, Theorem 3 continues to hold if we replace D with D_{x_0} in assumption (A6). Theorems 1 and 2 remain valid without any change, since:

$$p_{D_{x_0}}(x^* \mid z^*) \geqslant p_{D_{x_0}}(x_0 \mid z^*) \ \text{ and } \ p_{D_{x_0}}(x_0 \mid z^*) - p_D(x_0 \mid z^*) \geqslant p_{D_{x_0}}(x^* \mid z^*) - p_D(x^* \mid z^*)$$

together imply that $p_D(x^* \mid z^*) \geqslant p_D(x_0 \mid z^*)$, as required for expression (16) in our original proofs.

S2. Experiment Details and Additional Results

Automatic differentiation in TensorFlow is used to obtain gradients for both our revision procedure and the (stochastic) learning of neural network parameters. Throughout our applications, the GRU input is a vector-representation of each symbol in the sequence, taken from a dictionary of embeddings that is learned jointly with the neural network parameters via the Adam optimization algorithm of Kingma & Ba (2015). To ensure the decoder can actually generate variable-length sequences, a special <End> symbol is always included in $\mathcal{S}$ and appended at the end of each sequence in the training data. Note that all α-values stated in the text were actually first rescaled by $(2\pi)^{-d/2}$ before the REVISE procedure (to avoid confounding from the choice of latent-dimensionality d in the relationship between the listed α and characteristics of the resulting revisions).

S2.1. Simulation Study

When sampling a sequence for this simulation, we first draw its length uniformly from the range [10,20], and subsequently draw the symbols at each position following the probabilistic grammar of Table S1. Before its quality is evaluated, any proposed sequence whose length violates the [10,20] range is either truncated or extended via repeated duplication of the last symbol. In all models we apply, the encoder/decoder GRUs operate on input-embeddings of size 8, and the outcome-prediction model $\mathcal{F}$ is a feedforward network with one tanh hidden layer of size 128.

Rule	Probability
$s_t = A \mid s_{t-1} = C$	0.50
$s_t = B \mid s_{t-1} = A$	0.95
$s_t = D \mid s_{t-3} = D$	0.95
$s_t = E \mid s_{t-5} = E$	0.95
$s_t = J \mid s_{t-2} = H, s_{t-1} = I$	0.95
$s_t = I \mid s_{t-2} = I, s_{t-1} = H$	0.95
$s_t = B \mid s_{t-3} = B, s_{t-2} = C$	0.95
$s_t = F \mid s_{t-1} = F, t \geqslant 11$	0.95
$s_7 = G \mid s_6 = F$	0.95
$s_8 = G \mid s_7 = F$	0.50
$s_5 = C$	0.50
$s_{10} = C$	0.50
$s_{15} = C$	0.50
$s_{20} = C$	0.50

Table S1. Probabilistic grammar used to generate sequences $(s_1, \ldots, s_T)$ in our simulation. All events not listed here are assumed to occur randomly (uniformly among the remaining probability mass). When one or more conditioning statements are valid at a given t, we renormalize the probabilities for $s_t \mid s_1, \ldots, s_{t-1}$ before sampling the next character.

In the SEARCH procedure, evaluating 100 candidates took similar computation time as a typical run of our REVISE algorithm. Note that in this small scale simulation study, SEARCH is able to examine a nontrivial subset of the possible sequences around x_0. However, exponentially more randomly generated revisions would be needed to retain the performance of this SEARCH approach under longer sequences with larger vocabularies, whereas the computational complexity of our REVISE procedure scales linearly with such increases. Whereas the SEARCH method changes nearly every given initial sequence by a relatively similar amount, our REVISE procedure tends to either make larger changes or no change at all. As is desirable, our approach (particularly with adaptive decoding) tends to favor no change for x_0 where the corresponding latent posterior has high uncertainty, both because the VAE training objective urges all decodings in a large region around $E(x_0)$ to heavily favor x_0 and the invariance term $\mathcal{L}_{\text{inv}}$ encourages F to be more flat in such regions.

S2.2. Improving Sentence Positivity

For simplicity, our analysis of the beer reviews only considers sentences that are short ($\leqslant$ 30 words) and entirely composed of words that appear in $\geqslant 100$ other sentences. This restricts the size of the vocabulary to $|\mathcal{S}| \approx 5,500$. In this analysis, the SEARCH procedure is allowed to score 1000 candidate sequences, which is now far slower than our REVISE algorithm. In our models, GRUs $\mathcal{E}$ and $\mathcal{D}$ employ an embedding layer of size 128, the latent representations (and GRU hidden states h_t) have $d = 256$ dimensions, and $\mathcal{F}$ is feedforward network with one hidden layer of the same size (and tanh activations)

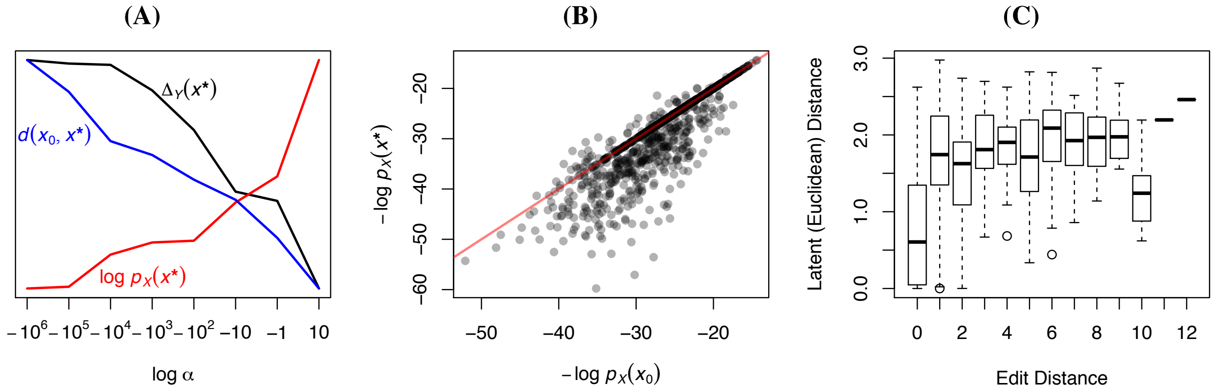

Figure S1. Behavior of the REVISE procedure in our simulation study. **(A)** Relationship between α and properties of revised sequence (averaged over same 1000 initial sequences $x_0 \sim p_X$, with units rescaled so that all curves share the same range): outcome improvement (black), edit distance (blue), marginal log-likelihood (red). **(B)** Likelihood of each original sequences vs. its revised version, when $\log \alpha = -10000$. The diagonal red line depicts the identity relationship $y = x$. **(C)** Boxplot of $||z^* - E(x_0)||_2$ values for each resulting value of $d(x_0, x^*)$ observed when $\log \alpha = -10000$. Note there were very few revisions where $d(x_0, x^*) > 8$.

followed by a sigmoid output layer. The language model L shares the same GRU architecture as our decoder network $\mathcal{D}$.

Examining the REVISE output, we find that punctuation patterns are quite often perfectly preserved in revisions (this is interesting since all punctuation characters are simply treated as elements of the vocabulary in the sequences). There exist many initialization-points where if unconstrained gradient ascent is run for a vast number of iterations with a large step-size, the resulting decoding produces the sentence: "excellent excellent excellent excellent excellent excellent excellent.", which is has near-optimal VADER sentiment but low marginal likelihood. Starting from other z-initializations, the decoding which results from a massive shift in the latent space often reverts to repetitions of a safe choice where each decoded word has high marginal likelihood, such as: "the the a the the the a the" or "tasting tasting tasting tasting tasting tasting tasting ".

S2.3. Revising Modern Text in the Language of Shakespeare

Sentences used in this analysis were taken either from the concatenated works of Shakespeare (Karpathy, 2015) or from various more contemporary texts (non-Shakespeare-authored works from the Brown, Reuters, Gutenberg, and FrameNet corpora in Python's NLTK library (Bird et al., 2009)). Here, we use the same architecture for networks $\mathcal{F}$, $\mathcal{E}$, $\mathcal{D}$ as in the previous beer-reviews application.

Model	Sentence	$\Delta_Y(x^*)$	$\Delta_L(x^*)$	$d(x^*, x_0)$
x_0	**caramel, fruit, sweetness, and a soft floral bitterness.**	-	-	-
$\log \alpha = -10000$	caramel, fresh, sweetness, quite soft and a good bitterness.	+1.88	-5.1	6
ADAPTIVE	caramel, fresh, sweetness, quite soft and a good bitterness.	+1.88	-5.1	6
$\log \alpha = -1$	caramel, fruit sweetness, and a soft floral nose.	+1.17	+0.2	1
$\lambda_{\text{inv}} = \lambda_{\text{pri}} = 0$	caramel, fruit sweetness, and a soft floral and tangy nose.	+1.17	-16.4	3
SEARCH	caramel, fruit sweetness, and a soft floral, cocoa.	+ 1.17	-7.0	2
x_0	**i like to support san diego beers.**	-	-	-
$\log \alpha = -10000$	i love to support craft beers!	+0.5	+1.6	4
ADAPTIVE	i like to support san diego beers.	0	0	0
$\log \alpha = -1$	i like to support craft beers!	+0.1	+2.6	3
$\lambda_{\text{inv}} = \lambda_{\text{pri}} = 0$	i like to support you know.	0	+3.7	3
SEARCH	i like to super support san diego.	+0.7	-2.9	2
x_0	**good carbonation makes for a smooth drinking experience.**	-	-	-
$\log \alpha = -10000$	good carbonation makes a great smooth drinking stuff.	+1.1	-1.1	3
ADAPTIVE	good carbonation makes a great smooth drinking stuff.	+1.1	-1.1	3
$\log \alpha = -1$	good carbonation makes for great smooth drinking.	+ 1.1	+3.0	2
$\lambda_{\text{inv}} = \lambda_{\text{pri}} = 0$	good carbonation makes for a smooth drinking like experience.	+0.7	-9.2	1
SEARCH	good carbonation makes for a drinking nice experience!	+0.9	-4.1	3
x_0	**i'm not sure how old the bottle is.**	-	-	-
$\log \alpha = -10000$	i definitely enjoy how old is the bottle is.	+3.0	-3.6	4
ADAPTIVE	i definitely enjoy how old is the bottle is.	+3.0	-3.6	4
$\log \alpha = -1$	i'm sure not sure how old the bottle is.	+2.5	-6.8	1
$\lambda_{\text{inv}} = \lambda_{\text{pri}} = 0$	i'm sure better is the highlights when cheers.	+3.3	-9.2	6
SEARCH	i 'm not sure how the bottle is love.	+2.3	-3.3	2
x_0	**what a great afternoon!**	-	-	-
$\log \alpha = -10000$	what a great afternoon!	0	0	0
ADAPTIVE	what a great afternoon!	0	0	0
$\log \alpha = -1$	what a great afternoon!	0	0	0
$\lambda_{\text{inv}} = \lambda_{\text{pri}} = 0$	what a great afternoon lace!	0	-8.2	1
SEARCH	what a solid great!	+0.19	-7.1	2
x_0	**the finish is a nice hoppy bitter, with ample spice.**	-	-	-
$\log \alpha = -10000$	the finish is a nice hoppy plant, with ample spice and great mouthfeel.	+2.5	-6.4	4
ADAPTIVE	the finish is a nice hoppy plant, with ample spice.	+1.3	-0.8	1
$\log \alpha = -1$	the finish is a nice hoppy plant, with ample spice.	+1.3	-0.8	1
$\lambda_{\text{inv}} = \lambda_{\text{pri}} = 0$	the finish is a nice hoppy bitter, with ample spice.	0	0	0
SEARCH	the finish is a nice hoppy bitter best, with ample spice.	+2.0	-7.9	1

Table S2. Additional examples of held-out beer reviews x_0 (in bold) revised to improve their VADER sentiment. Underneath each sentence, we show the revision produced by each different method along with the true outcome improvement $\Delta_Y(x^*) = \mathbb{E}[Y \mid X = x^*] - \mathbb{E}[Y \mid X = x_0]$ (rescaled by the standard deviation of outcomes in the training data), estimated change in marginal likelihood $\Delta_L(x^*) = \log L(x^*) - \log L(x_0)$, and Levenshtein (edit) distance $d(x^*, x_0)$.

# Steps	Sentence
x_0	**you find the evidence of that in the chart on this page.**
100	you find the evidence of that in the chart on this page.
1000	you find the chart of action in this page.
5000	you find the chart of the chart that page of action in this page.
10000	find you in this page of the way of your highness.
x^*	you speak of the chart in this page of the lord.
x_0	**somewhere, somebody is bound to love us.**
100	somewhere, somebody is bound to love us.
1000	courage, honey, somebody is bound to love us!
5000	courage man; 'tis love that is lost to us.
10000	thou, within courage to brush and such us brush.
x^*	courage man; somebody is bound to love us.
x_0	**the story of the fatal crash is not fully known**
100	the story of the injured is not known.
1000	the story of our virtue is not yet known.
5000	the story of our virtue is not given me yet.
10000	the virtue of our story is not yet.
x^*	the story of our virtue is not yet known.
x_0	**this is the root issue for which the united states should stand.**
100	this is the root issue which is an issue on the united states.
1000	the root issue is that the dialogue itself should stand provided.
5000	the general is for the root chief held for which is thy tale.
10000	this the shallow is sworn thee. shallow for thee.
x^*	the root issue is the national dialogue from thine.
x_0	**there is no such magic in man-made laws.**
100	there is no such magic of man in such magic.
1000	there is no magic of man in such magic.
5000	there is no magic question with such a man in man.
10000	there is no magic in revolution and made no such india.
x^*	there is no magic in such noble birth;
x_0	**check the quality of the water.**
100	check the quality of the water.
1000	check the quality of thy water.
5000	check the quality of thy quality.
10000	check the king of gloucester.
x^*	check the quality of thy water.
x_0	**what are you doing here?**
100	what are you doing here?
1000	what are you doing here?
5000	cardinal what does thou live here?
10000	cardinal what does thou live here?
x^*	does thou live here?

Table S3. Adaptive decoding from various latent Z configurations encountered at the indicated number of (unconstrained) gradient steps from $E(x_0)$, for the model trained to distinguish sentences from Shakespeare vs. contemporary authors. Shown first and last are the initial sequence x_0 and the revision x^* returned by our REVISION procedure (constrained with $\log \alpha = -10000$).

S3. Proofs and Auxiliary Lemmas

Proof of Theorem 1.

By the definition of x^*, we have:

$$
\begin{aligned}
& p_D(x^* \mid z^*) \geqslant p_D(x_0 \mid z^*) && (16)\\
\implies & p_X(x^*) \geqslant \frac{p(z^* \mid x_0)}{p(z^* \mid x^*)} \cdot p_X(x_0) && \text{by Bayes' rule}\\
& \geqslant \frac{\gamma q_E(z^* \mid x_0)}{p(z^* \mid x^*)} \cdot p_X(x_0) \text{ with probability} \geqslant 1-\delta
\end{aligned}
$$

by assumptions (A1) and (A2) combined via the union bound. Finally, from the definitions in REVISE, we have that $z^* \in \mathcal{C}_{x_0}$, which implies $q_E(z^* \mid x^*) \geqslant \alpha$. □

Lemma 1. *If (A1) holds, then for z^* defined in* REVISE*: $z^* \in B_R(0)$ with probability $\geqslant 1 - \frac{\delta}{2}$ (over $x_0 \sim p_X$).*

Proof. Recall that $B_R(0)$ is defined as the Euclidean ball of radius R centered around 0. We show:

$$||z^* - E(x_0)|| \leqslant \frac{1}{2}R \tag{17}$$

and with probability $\geqslant 1 - \frac{\delta}{2}$:

$$||E(x_0)|| \leqslant \frac{1}{2}R \tag{18}$$

Subsequently, the triangle inequality completes the proof.
To prove (17), we recall that from our definition in (3): $q_E(z \mid x_0)$ is a Gaussian distribution with mean $E(x_0)$ and diagonal covariance $\Sigma_{z|x}$ where each entry is $\leqslant 1$. Furthermore, the definitions in REVISE ensure $z^* \in \mathcal{C}_{x_0} \implies q(z^* \mid x_0) \geqslant \alpha$. Defining $K = -2\log[(2\pi)^{d/2}|\Sigma_{z|x}|^{1/2}\alpha]$ which specifies the level-α isocontour of the $N(0, \Sigma_{z|x})$ density, we have:

$$
\begin{aligned}
& q(z^* \mid E(x_0) \geqslant \alpha\\
\implies & (z^* - E(x_0))^T \Sigma_{z|x}^{-1} (z^* - E(x_0)) \leqslant K\\
\implies & ||z^* - E(x_0)|| \leqslant \sqrt{K \cdot \lambda_{\max}(\Sigma_{z|x})} \leqslant \frac{1}{2}R_1
\end{aligned}
$$

where $\lambda_{\max}(\Sigma_{z|x})$ is the largest eigenvalue of $\Sigma_{z|x}$ and $\lambda_{\max}(\Sigma_{z|x}) \leqslant 1,\ |\Sigma_{z|x}|^{1/2} \leqslant 1$ for our $q_E(z \mid x)$.

Now, define $\mathcal{R} = \{x \in \mathcal{X} : E(x) > \frac{1}{2}R\}$, and let $\widetilde{Z} \sim q_Z$ as defined in (10). To prove (18), we note that for all $x \in \mathcal{R}$: $q_E(z \mid x)$ is a diagonal Gaussian distribution centered around $E(x)$ which has norm $> R/2$. Thus:

$$
\begin{aligned}
\frac{\gamma}{4} \cdot p_X(\mathcal{R}) <& \gamma \sum_{x \in \mathcal{R}} \int_{||z|| \geqslant \frac{1}{2}R} q_E(z \mid x) \, \mathrm{d}z \, p(x) = \gamma \cdot \Pr\left(||\widetilde{Z}|| \geqslant \frac{1}{2}R\right)\\
\leqslant & \Pr\left(||Z|| \geqslant \frac{1}{2}R\right) && \text{by the second condition in (A1)}\\
\leqslant & \Pr\left(||Z|| \geqslant \frac{1}{2}R_2\right) && \text{as we defined } R \geqslant R_2
\end{aligned}
$$

Since $Z \sim N(0, \mathbf{I})$ under our prior, $||Z||^2 \sim \chi^2_d$.
Applying the Chernoff bound to the tail of the χ^2 distribution (Dasgupta & Gupta, 2002), we thus obtain:

$$\Pr\left(||Z||^2 \geqslant \frac{1}{4}R_2^2\right) \leqslant \left[\frac{1}{4}R_2^2 \cdot \exp\left(1 - \frac{1}{4}R_2^2\right)\right]^{d/2} \leqslant \left[\exp\left(1 - \frac{1}{16}R_2^2\right)\right]^{d/2}$$

which implies $p_X(\mathcal{R}) < \delta/2$ by our definition of R_2. □

Proof of Theorem 2.

For $\epsilon \in (0,1]$, let $B_\epsilon(z)$ denote the ϵ-ball centered at z. We have:

$$p_X(x^*) = \int p_D(x^* \mid z) p_Z(z) \, \mathrm{d}z$$

$$\geqslant \Pr(Z \in B_\epsilon(z^*)) \left[p_D(x^* \mid z^*) - L\epsilon \right]$$

assuming $z^* \in B_R(0)$, which occurs with probability $\geqslant 1 - \delta/2$ by Lemma 1

$$\geqslant \Pr(Z \in B_\epsilon(z^*)) \left[p_D(x_0 \mid z^*) - L\epsilon \right] \qquad \text{by (16)}$$

$$= \Pr(Z \in B_\epsilon(z^*)) \left[\frac{p(z^* \mid x_0)}{p_Z(z^*)} p_X(x_0) - L\epsilon \right]$$

$$\geqslant \Pr(Z \in B_\epsilon(z^*)) \left[\gamma \frac{q_E(z^* \mid x_0)}{p_Z(z^*)} p_X(x_0) - L\epsilon \right]$$

assuming $z^* \in B_R(0)$ and x_0 satisfies the (A1) inequality, which occurs with probability $\geqslant 1 - \delta$ by the union bound

$$\geqslant \frac{\Pr(Z \in B_\epsilon(z^*))}{p_Z(z^*)} \left[\gamma \alpha p_X(x_0) - L\epsilon \right] \qquad \text{since } p_Z(z^*) < 1 \text{ and } z^* \in \mathcal{C}_{x_0} \implies q_E(z^* \mid x_0) \geqslant \alpha$$

$$\geqslant \frac{\min\limits_{||\Delta||=\epsilon} p_Z(z^* + \Delta)}{p_Z(z^*)} \mathrm{Vol}(B_\epsilon(z^*)) \left[\gamma \alpha p_X(x_0) - L\epsilon \right] \qquad \text{where Vol}(\cdot) \text{ denotes the Lebesgue measure}$$

$$\geqslant \exp\left(-\frac{1}{2} \left[||z^*|| \epsilon + \epsilon^2 \right] \right) \mathrm{Vol}(B_\epsilon(z^*)) \left[\gamma \alpha p_X(x_0) - L\epsilon \right]$$

by exploiting the fact that $p_Z = N(0, I)$ and subsequent application of the Cauchy-Schwarz inequality

$$\geqslant \exp\left(-\frac{||z^*|| + 1}{2} \right) \cdot \mathrm{Vol}(B_\epsilon(z^*)) \cdot \left[\gamma \alpha p_X(x_0) - L\epsilon \right] \qquad \text{for any } \epsilon \in (0,1]$$

$$\geqslant \exp\left(-\frac{R + 1}{2} \right) \cdot \mathrm{Vol}(B_\epsilon(z^*)) \cdot \left[\gamma \alpha p_X(x_0) - L\epsilon \right] \qquad \text{since we already assumed } z^* \in B_R(0).$$

We conclude the proof by selecting $\epsilon = \frac{\gamma\alpha(d+1)}{L(d+2)} p_X(x_0)$ which maximizes the lower bound given above. □

Proof of Theorem 3.

Suppose for $x_0 \in \mathcal{R}$, the corresponding revision $x^* \notin \mathcal{E}$. Then:

$$\begin{aligned} \Pr(X \in \mathcal{E} \cap \mathcal{R}) &\leqslant 1 - p_X(x^*) - \Pr(X \in \mathcal{E} \backslash \mathcal{R}) \\ &\leqslant 1 - \kappa - \Pr(X \in \mathcal{E} \backslash \mathcal{R}) \end{aligned}$$

Since (A5) implies $\Pr(X \in \mathcal{E}^C) < \kappa$, we also have:

$$\begin{aligned} \Pr(X \in \mathcal{E} \cap \mathcal{R}) &= 1 - \Pr(X \in \mathcal{E}^C) - \Pr(X \in \mathcal{E} \backslash \mathcal{R}) \\ &> 1 - \kappa - \Pr(X \in \mathcal{E} \backslash \mathcal{R}) \end{aligned}$$

which is a contradiction. Thus, we must have $x^* \in \mathcal{E}$ if $x_0 \in \mathcal{R}$, which occurs with probability $\geqslant 1 - \delta/2$.

Lemma 1 ensures that under (A1): $z^* \in B_R(0)$ with probability $\geqslant 1 - \delta/2$, implying $|F(z^*) - F(E(D(z^*)))| \leqslant \epsilon_{\text{inv}}$ with the same probability. Consequently, we have:

$$\begin{aligned} F(z^*)) - F(E(x_0)) &\leqslant F(E(D(z^*)) - F(E(x_0)) + \epsilon_{\text{inv}} && \text{with probability} \geqslant 1 - \tfrac{\delta}{2} \\ &\leqslant F(E(x^*)) - \mathbb{E}[Y \mid X = x_0] + \epsilon_{\text{inv}} + \epsilon_{\text{mse}} && \text{with probability} \geqslant 1 - \tfrac{\delta}{2} - \kappa \text{ by the union bound} \\ &\leqslant \mathbb{E}[Y \mid X = x^*] - \mathbb{E}[Y \mid X = x_0] + \epsilon_{\text{inv}} + 2\epsilon_{\text{mse}} && \text{with probability} \geqslant 1 - \tfrac{\delta}{2} - \kappa - \tfrac{\delta}{2} \text{ by the union bound} \end{aligned}$$

The inequality in the other direction is proved via similar reasoning. □

Additional References for the Supplementary Material

Bird, S., Klein, E., and Loper, E. *Natural Language Processing with Python*. O'Reilly Media, 2009.

Dasgupta, S. D. A. and Gupta, A. K. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures and Algorithms*, 22:60–65, 2002.

Karpathy, A. The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy blog*, 2015. URL `karpathy.github.io`.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.